

# International Journal of Big Data and Analytics in Healthcare

Volume 6 • Issue 2 • July-December 2021 • ISSN: 2379-738X • eISSN: 2379-7371



**IGI PUBLISHING**  
AN IMPRINT OF IGI GLOBAL  
[WWW.IGI-GLOBAL.COM](http://WWW.IGI-GLOBAL.COM)

## EDITOR-IN-CHIEF

Mu-Yen Chen, National Cheng Kung University, Taiwan

## ASSOCIATE EDITORS

Amir Rahmani, University of California Irvine, USA

Arun Kumar Sangaiah, Vellore Institute of Technology, India

Edwin Lughofer, University of Linz, Austria

Guillermo Lopez Campos, Queen's University Belfast, Australia

Hsin-Te Wu, National Ilan University, Taiwan

Hsiu-Sen Chiang, National Taichung University of Science and Technology, Taiwan

Jong Hyuk Park, Seoul National University of Science and Technology, South Korea

Karin Verspoor, University of Melbourne, Australia

Min Chen, Wenzhou University, China

Neil Y. Yen, University of Aizu, Japan

Parameshachari B.D., GSSS Institute of Engineering and Technology for Women, Mysuru, India

Sugam Sharma, Iowa State University, USA

Tien-Chi Huang, National Taichung University of Science and Technology, Taiwan

Yeliz Karaca, University of Massachusetts Medical School, USA

Yung-Kuan Chan, National Chung Hsing University, Taiwan

## EDITORIAL REVIEW BOARD

Abdelhamid Bouchachia, Bournemouth University, UK

Alexandra Gorelik, University of Melbourne, Australia

Ching-Ta Lu, Asia University, Taiwan

Frank Lee, University of Guam, USA

Haiyan Su, Montclair State University, USA

Hoda Moghimi, RMIT University, Epworth Healthcare, Australia

Jason C. Hung, The Overseas Chinese Institute of Technology, Taiwan

Jeng-Wei Lin, Tunghai University, Taiwan

Jung-wen Lo, National Taichung University of Science and Technology, Taiwan

Jun-Hong Shen, Asia University, Taiwan

Jyotir Moy Chatterjee, Lord Buddha Education Foundation, Nepal

M. Rathan, REVA University, India

Mamoon Rashid, Lovely Professional University, India

Nikolaos Korfiatis, University of East Anglia, UK

Paul Woolman, NHS Scotland, UK

Pedro Peris-Lopez, Universidad Carlos III de Madrid, Spain

Rada Hussein, Information Technology Institute, USA

Vanessa Aguiar-Pulido, Weill Cornell Medicine, USA

Wadee S. Alhalabi, Effat University, Saudi Arabia

Xiong Li, Hunan University of Science and Technology, China

# Call for Articles

## International Journal of Big Data and Analytics in Healthcare

Volume 6 • Issue 2 • July-December 2021 • ISSN: 2379-738X • eISSN: 2379-7371

### MISSION

The mission of the **International Journal of Big Data and Analytics in Healthcare (IJBD AH)** is to provide timely and innovative research on the ways in which big data is revolutionizing the medical and healthcare fields. This journal aims to encourage the further development of applications and practice relating to the management and analysis of large amounts of data in the healthcare sector as well as provide a framework for future research in the field.

### COVERAGE AND MAJOR TOPICS

**The topics of interest in this journal include, but are not limited to:**

Ambient Intelligence • Artificial Intelligence • Behavioral and clinical data • Big data for disease diagnosis • Big data for electronic health records • Body sensor networks • Cloud Computing • Context-aware and emotion-aware service • Data mining and data stream mining • Data Visualization • Deep Learning • Embedded system and software • Internet of Things (IoT) • Medical image processing and pattern recognition • Security, Privacy and Trust • Soft Computing • Text Mining • Ubiquitous Computing • Wearable Computing

**ALL INQUIRIES REGARDING IJBD AH SHOULD BE DIRECTED TO THE ATTENTION OF:**

Mu-Yen Chen, Editor-in-Chief • [IJBD AH@igi-global.com](mailto:IJBD AH@igi-global.com)

**ALL MANUSCRIPT SUBMISSIONS TO IJBD AH SHOULD BE SENT THROUGH THE ONLINE SUBMISSION SYSTEM:**

<http://www.igi-global.com/authorseditors/titlesubmission/newproject.aspx>

IDEAS FOR SPECIAL THEME ISSUES MAY BE SUBMITTED TO THE EDITOR(S)-IN-CHIEF

**PLEASE RECOMMEND THIS PUBLICATION TO YOUR LIBRARIAN**

For a convenient easy-to-use library recommendation form, please visit:

<http://www.igi-global.com/IJBD AH>



# InfoSci®-Journals

A Database of Over 25,000+ Articles With Over 1,000,000+  
Citation References Sourced From 185+ Scholarly Journals



GAIN ACCESS TO **HUNDREDS OF**  
SCHOLARLY JOURNALS AT **A FRACTION**  
OF THEIR INDIVIDUAL LIST **PRICE**.

## InfoSci®-Journals Database

**InfoSci®Journals** database is a collection of over 185+ scholarly journals, encompassing groundbreaking research from prominent experts worldwide that spans over 350+ topics in 11 core subject areas including business, computer science, education, science and engineering, social sciences, and more. With all of the journals featured in prestigious indices including Web of Science® and Scopus®, this database option allows libraries to subscribe to the entire journal collection priced at the cost of just a few single-title subscriptions.

### Open Access Fee Waiver (Offset Model) Initiative

For any library that invests in IGI Global's InfoSci-Journals database and/or a subset of this research database, IGI Global will match the library's investment with a fund of equal value to go toward subsidizing the OA article processing charges (APCs) for their students, faculty, and staff at that institution when their work is submitted and accepted under OA into an IGI Global journal.\*

### INFOSCI® PLATFORM FEATURES

- No DRM
- No Set-Up or Maintenance Fees
- A Guarantee of No More Than a 5% Annual Increase
- Full-Text HTML and PDF Viewing Options
- Downloadable MARC Records
- Unlimited Simultaneous Access
- COUNTER 5 Compliant Reports
- Formatted Citations With Ability to Export to RefWorks and EasyBib
- No Embargo of Content (Research is Available Months in Advance of the Print Release)

\*The fund will be offered on an annual basis and expire at the end of the subscription period. The fund would renew as the subscription is renewed for each year thereafter. The open access fees will be waived after the student, faculty, or staff's paper has been vetted and accepted into an IGI Global journal and the fund can only be used toward publishing OA in an IGI Global journal. Libraries in developing countries will have the match on their investment doubled.



To Learn More or To Purchase This Database:

[www.igi-global.com/infosci-journals](http://www.igi-global.com/infosci-journals)

eresources@igi-global.com • Toll Free: 1-866-342-6657 ext. 100 • Phone: 717-533-8845 x100

# A New Internet Public Opinion Evaluation Model: A Case Study of Public Opinions on COVID-19 in Taiwan

Sheng-Tsung Tu, Ming Chuan University, Taiwan

Louis Y. Y. Lu, Yuan Ze University, Taiwan

Chih-Hung Hsieh, Yuan Ze University, Taiwan

Chia-Yu Wu, Yuan Ze University, Taiwan

## ABSTRACT

This research retrieved public opinions on the novel coronavirus pandemic with the aid of the DiVoMiner. The data were collected by setting keywords via qualitative comparative analysis (QCA) and automated computational approach, and the collected data were analyzed subsequently. The present study divided keyword collections into three categories, namely the name of diseases, government policies, and COVID-19 events. It was found the retrieved internet public opinions on COVID-19 were the largest in number and contained the least noise when the three categories of keywords appeared at the same time. Therefore, the data of internet public opinions = the name of diseases  $\times$  (government policies + COVID-19 events). This research found that an event that happens daily will affect the number of internet public opinions on social media and forums after it has been reported. The strong negative emotion conveyed through the internet public opinion may turn into a positive one if the event is dealt with properly after positive focus words represent the same proportion as negative ones.

## KEYWORDS

Automated Computational Approach, COVID-19, Divominer, Internet Public Opinion, Qualitative Comparative Analysis

## 1. INTRODUCTION

In December 2019, cluster infection of unidentified COVID-19 occurred. A novel coronavirus, which can spread across species and through human-to-human contact, was isolated and sequenced, which was named “severe acute respiratory syndrome coronavirus 2 (i.e. SARS-CoV-2). On February 11, 2020, the World Health Organization (WHO) officially named the severe infectious pneumonia caused by this virus Coronavirus Disease-19 (i.e. COVID-19). Due to its rapid spread and inevitable high infectivity, it has been classified as a major infectious disease by WHO. It has spread to 185 countries and regions (Di Gennaro et al., 2020). As of November 10, 2020, there were a total of 50,913,451

DOI: 10.4018/IJBDAH.287603

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.



confirmed cases worldwide, resulting in 1,263,089 deaths according to the COVID-19 dashboard of Johns Hopkins University.

In the modern society that features information explosion, people are becoming increasingly dependent on the Internet. Against this background, the young are more willing to voice their opinions on the Internet, commenting on news, responding to the posts on social media, and participating in discussion on a forum. Thus, their remarks about an event can be regarded as Internet public opinions. Internet users' increasing adherence to the Internet enables the authorities to promote their policies via social media and traditional news media to regard social media like Facebook and Youtube as a platform to disseminate information.

The Taiwanese authorities took the lead in COVID-19 prevention because of an article posted on PTT Gossiping. On December 31<sup>st</sup>, 2019, the Internet user “nomorepipe” published the article “Suspected SARS Coronavirus Cluster Infection Broke out in Wuhan?” The netizen posted a test report concerning the coronavirus by Li Wenliang, who is claimed Coronavirus Whistleblower doctor. The article led to heated discussion, which drew the attention of the Centers for Disease Control and Prevention (CDC). Therefore, Taiwan initiated its epidemic prevention in January 1, 2020, which was earlier than other countries and regions across the world.

In brief, the present study aimed to use DiVoMiner to retrieve the data about Internet public opinions on COVID-19 between December 31, 2019, the time when public opinions on COVID-19 first appeared, and June 7, 2020, the time when Central Epidemic Command Center (CECC) “unsealed” Taiwan. Afterward, QCA and ACA were employed to set the keywords that were used to retrieve and analyze data, expecting to clarify the following questions:

1. How many reports and discussions about COVID-19 are there on the Internet, including news media, social media and forums?
2. What are the words that Internet users choose to comment on COVID-19 events?
3. How do the emotions that the Internet public opinions on COVID-19 convey change?
4. Is there a correlation between the numbers of public opinions about COVID-19 on different media?

## **2. INTERNET PUBLIC OPINIONS**

Compared with traditional media, the Internet has fewer limitations caused by carriers. Moreover, the Internet gradually changes the way of information transmission: the information was previously disseminated from “one to one” and “one to many”, but now it is transmitted from “many to many”. On the other hand, the information flow on the Internet spreads fast but gets insufficient management. As for public opinions, they are complicated and may be antagonistic to each other. Worse still, readers' subjective judgment during information dissemination makes the media unable to play the role of “gatekeeper” like before. As a consequence, the authenticity of information on the Internet cannot be guaranteed. In this context, a great deal of ever-flowing information inevitably includes false and unconfirmed one. Therefore, opinion leaders may take Internet users to a wrong direction, and some even deliberately take public opinions to a direction that helps them achieve their goals (Huang Wei, Li Rui, & Meng Jialin, 2015).

Previous researches on “Internet public opinions” have defined it. Overall, Internet public opinions can be deemed as the emotions, ideas, opinions, attitudes and social influence that a netizen<sup>1</sup> has toward a social problem, public event, ideology, and morality with the Internet as the carrier and an event as the core. The rapid spread of Internet public opinions exerts a great impact on many aspects of social life. With the rapid development of the Internet, Internet public opinions form quickly and its social influence is getting increasingly larger. The influence of Internet media has far exceeded that of traditional media, such as newspapers, radio and television, which makes it hard for traditional media to continue its development. Therefore, traditional media has been undergoing adjustments.

As well, Internet public opinions change its way of presentation rapidly. In the initial stage, they were mainly manifested in news reviews, PTT forums, comments, and reposts. They are now mainly presented via Facebook and Instagram.

In fact, public opinions on the Internet are similar to that in real world, and the remarks and deeds of opinion makers on the Internet and in real world are both similar and different. In reality, residents can be classified into different classes, social groups or interest groups. In the Internet society, netizens participate in different online communities according to their own interests, preferences, and values. The difference between the two lies in the fact that in the Internet society, netizens have no labels that they bear in real world. No matter what role they play in real world, they can equally express their opinions on the Internet in an anonymous manner. The equality between communication subjects on the Internet also brings to an end the era of discourse monopoly between the government and the media. Due to the obstacles from society as well as cultural and ideological influence, people are often unable to express their emotion, wish, dissatisfaction and anger in the real world. The anonymity of netizens on the Internet provides them with the opportunity to express the dissatisfaction that they have experienced in the real world. This way, Internet public opinions come into being. Moreover, some Internet public opinions will affect the real society, causing the reaction of social subjects, such as Mass Protest over Corporal Chung-Chiu Hung, Jasmine Revolution, the Arab Spring events, etc.

The literature review reveals multiple problems as follows:

1. **Data from Limited Number of Websites:** Some of the previous researches on big data often wrote web crawlers through word patterns and programs to retrieve data from specific websites. This way, relevant data only come from given websites, which may cause research inaccuracy and loss of reliability and validity unless a research aims to measure the public opinion of given websites or Internet forums.
2. **Problems about Setting Keywords:** Previous researches in this field often used a single keyword or keyword combination (usually four to five keywords) to retrieve data about public opinions. However, if researchers set keywords this way, they tend to obtain noise data, which necessitates time-consuming data collation. Research on Internet public opinions values efficiency, while data collation may lead to the loss of timeliness that public opinion evaluation requires.
3. **Excessive Intervention:** At the initial stage of the evaluation of Internet public opinions, trained coders conducted data coding and emotional judgment in some researches. Compared with computer-based calculation, manual judgment unavoidably causes judgmental inaccuracies, thus resulting in judgmental errors. In this respect, the computer-based calculation has greater chance to improve the overall accuracy of Internet public opinion evaluation.

### 3. METHODOLOGY

#### 3.1 Computer-Based Calculation

Natural language processing (NLP) of machine learning enables computers to automatically analyze massive data, including trend analysis, emotional analysis, and breaking-down-sentence analysis. Therefore, this research employed DiVoMiner to structure the data of different sources through rigorous analysis and a reliability monitoring mechanism. This way, all the data were gathered in a single platform for internal auditing and filtration, after which rigorous content analysis method, which involves real-time coding, examination, monitoring and presentation, was utilized to visualize the data and conduct valuable semantic analysis at the same time. This way, this research obtained a report that has insights and facilitates decision-making. At present, DiVoMiner collects data from major news platforms, forums, PTT, Facebook, Instagram, Youtube, etc.

The semantic machine learning model having been introduced into DiVoMiner, the textual mining and analysis platform can not only be used for manual coding, but also for machine learning coding,

multidimensional analysis, analysis of the correlation between multiple variables, cross analysis, regression analysis, statistical verification, and the creation of word clouds. DiVoMiner also has a complete mechanism that monitors coding performance, which enables it to control the efficiency and accuracy of sampling. In addition, DiVoMiner adopts different inferential statistical methods to verify the representativeness of the results.

### 3.2 Keyword Setting

In the present study, keyword combinations consisted of three categories, i.e. the name of diseases, government policies, and COVID-19 events, which were analyzed via QCA.

QCA was a method published by the American social science scholar Charles C. Ragin in 1984. It was first proposed in 1987 (Rihoux, 2003; Dixon-Woods et al., 2005; Rihoux, 2006; Schneider & Wagemann, 2006), but it was not widely used until 1997 (Ragin, Shulman, Weinberg, & Gran, 2003). QCA, a method that integrates quantitative and qualitative research, features an analytical model of set theory. QCA is suitable for analyzing small- and medium-sized samples, while it has been used to analyze large-sized samples in a small number of studies; it has been compared with quantitative analysis like logistic regression, and the results obtained are almost the same (Grendstad, 2007).

QCA creates a truth table based on the dichotomy of “0” and “1” for the topic-related factors. The appearance of the code 0 means the absence of representative factors whereas the appearance of the code 1 indicates the existence of the representative factors. Afterward, Boolean Logic was used to figure out the configurations (Ragin, 1987), which provides a relatively objective basis for the explanation of cause-effect relations (i.e. causation).

When three factors are used for QCA, there are eight ( $2^3 = 8$ ) combinations. For instance, X is a dependent variable, a, b, and c are independent variables; ab represents the combinations that a and b appear simultaneously, while ac represents the combinations that a and c appear at the same time. Therefore,  $X = ab + ac$  indicates that X is the connected set of the combinations ab plus the combinations ac. Meanwhile, it means that the dependent variable X will have the maximum, so the equation can also be represented as  $X = a * (b + c)$ , indicating that a is the necessary factor for X (Ragin, 1999a; Ragin, 1999b; Rihoux, 2006). Thus, it means that the appearance of the factor a inevitably will result in the emergence of the dependent variable X.

QCA is a qualitative research method, which necessitates full discussion based on relevant data when researchers select factors and conduct coding. Moreover, whether the obtained results are proper should be discussed as well. An advantage of QCA is that it provides a systematic analysis of complicated and massive qualitative data. In addition, researchers obtain consistent equation as long as the researchers use the same variable for analysis and the same coding. Therefore, QCA has the characteristics of universal applicability and extrapolation, and reaches a level that qualitative research could not achieve before (Rihoux, 2003; Rihoux, 2006). Moreover, QCA can not only verify regular combinations found in existing researches, but also find out accidental or abnormal factors.

Regarding the name of diseases, keyword setting experienced two stages. First, academic terms related to Wuhan pneumonia that appeared in relevant reports both home and abroad were measured, such as Wuhan pneumonia, novel coronavirus, nCoV, novel coronavirus 2019, 2019-nCoV, severe infectious pneumonia, novel coronavirus pneumonia, NCP, COVID-19, and novel corona pneumonia. Subsequently, the keywords used for the first measurement were modified, after which they were discussed by experts and scholars of the related fields. The keywords were finally set after data cleaning.

The keywords obtained after data cleaning were as follows:

*Wuhan pneumonia or novel coronavirus or nCoV or 2019 novel coronavirus or 2019-nCoV or severe special infectious pneumonia or novel coronavirus pneumonia or NCP or COVID-19 or novel corona pneumonia or MERS or severe special infectious pneumonia.*

As for the keyword combinations of government policies, they were set based on the categories and keywords listed on the COVID-19 epidemic prevention network established by the CECC. The keywords were as follows:

1. **Community-based Epidemic Prevention:** Home quarantine or delaying the start of school or epidemic prevention care leave or autonomous health management or travel history or contact history or centralized quarantine or large-scale rally or “epidemic prevention” hotel or “epidemic prevention” taxi or social distancing or body temperature measurement or wearing masks or closing business or control of the flow of people or new life under “epidemic prevention” or real-name system or real-name registration.
2. **Border policy:** Boarding quarantine or travel epidemic or Arrival Health Declaration Form or ban on entry or visa control or flight dedicated to Taiwanese businessmen or entry ban or prohibited entry or transfer prohibited or Taiwan-bound flight or opening dedicated flights or monitoring of the flow of people or negative or airport inspection or airport quarantine or Chinese nationality or conditional access to Taiwan or overseas students.
3. **Supplies:** Epidemic prevention items or restricted exports or Mask National Team or masks or Mask Real-Name System or alcohol or mask map or real-name masks or mask 2.0 or eMask the Name-based System for Mask Purchasing or opening exports.
4. **Relief or compensation policy:** Special Act for Prevention, Relief and Revitalization Measures for Severe Pneumonia with Novel Pathogens or relief or epidemic prevention compensation or relief scheme or living allowance.
5. **Control of medical institutions:** Restrictions from going abroad or group gatherings or access control or personnel control or easing control.
6. **Inspection / Research & Development (R & D) Policy:** Remdesivir or vaccine or quarantine or inspection or self-paid examination.

The keyword combinations for COVID-19 events were set based on relevant news media reports. The keyword combinations that appear in aforementioned categories will not be included in this category, such as masks, relief policies, etc.

1. **Delaying the Start of School:** (Senior high school and below and delaying the start of school) or (colleges and delaying the start of school) or (Students from Chinese mainland and delaying the arrival in Taiwan) or (Students from Hong Kong and Macau and suspending entries) or epidemic prevention care leave or soothing schooling program.
2. **Flight dedicated to Taiwanese businessmen:** Flight dedicated to Taiwanese businessmen in Wuhan or Flight dedicated to Taiwanese businessmen or evacuation flights or evacuation flights from Wuhan.
3. **Religious activities:** Mazu or Baishatun Mazu or Dajiamia or Dajia Mazu or Baishatun Gongtian Palace or Dajia Zhenlan Palace or Pilgrimage or Mazu circumnavigation activity.
4. **Hoarding:** Hoarding or rushing to purchase goods or panic buying or storing up food or replenishing or hoarding masks or hoarding food or rushing to buy toilet paper or toilet paper.
5. **Dunmu fleet infection:** Dunmu fleet or confirmed cases on a navy warship or Panshi Fleet.

The above three keyword combinations were used to make a truth table through QCA, as shown in Table 1.

In the present research, QCA was conducted. It was found that X, the data about Internet public opinions on COVID-19, can be calculated by the following equation:

$$X = \text{the name of diseases} \times (\text{government policies} + \text{COVID-19 events})$$

**Table 1. A truth table for covid-19 keyword combinations**

Code	1	2	3	4	5	6	7	8
Name of diseases	0	1	1	1	1	0	0	0
government policies	0	1	1	0	0	1	1	0
COVID-19 events	0	1	0	1	0	1	0	1

Source: collated by the present study

Consequently, the keywords used to retrieve data include two categories; one is the name of diseases while the other is about government policies and COVID-19 events. After the keyword combinations were confirmed, the evaluation of the present study was conducted based on the keyword combinations.

## 4. RESEARCH RESULTS

This study retrieved 1,696,010 articles related to COVID-19 from the DiVoMiner database of Internet public opinions. The first article, titled “Suspected SARS coronavirus cluster infection broke out in Wuhan?”, appeared on PTT gossiping on December 31<sup>st</sup>, 2019. The Internet user named nomorepipe posted a test report concerning the novel coronavirus by Li Wenliang, who is claimed Coronavirus Whistleblower doctor. The article resulted in heated discussion, which attracted the attention of Taiwan’s Centers for Disease Control and Prevention (CDC). As a result, Taiwan initiated its epidemic prevention on January 1, 2020, which was earlier than other countries and regions across the world. Despite the effort, Taiwan had the first confirmed COVID-19 case, which was overseas imported, on January 21, 2020, and the first local COVID-19 case on January 28, 2020.

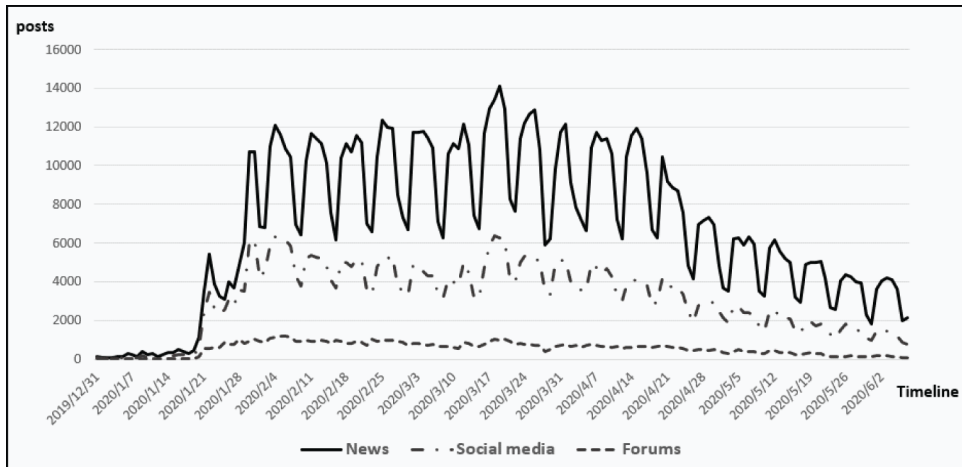
Among the 1,680,899 articles related to COVID-19, 65.09% were published on news media ( $n = 1,094,044$ ), 29.66% came from social media ( $n = 498,562$ ), and 5.25% from forums and discussion boards ( $n = 88,293$ ). Of 498,562 articles from social media, a dominant majority were published on Facebook (91.49%;  $n = 456,113$ ), and merely 8.51% came from Youtube and Intargram ( $n = 45,449$ ). Regarding news media, an average of 6,838 reports and discussions about COVID-19 (range 0-14,088) were published, while on social media, there were 3,116 posts, reposts, and discussions about COVID-19 (range 0-6,359). As for forums and discussion boards, an average of 552 articles and discussions about COVID-19 (range 0-1,210).

Regarding the number of Internet public opinions on COVID-19, as of January 19, the number of public opinions was less than 1,000 each day. However, Taiwan expanded its epidemic prevention area to airports on January 19 because it had 4 suspected cases on January 17, 2020. Since January 20, 2020, the number of Internet public opinions has exceeded 1,000 each day. Therefore, January 20 can be regarded as the starting point for the drastic growth of Internet public opinions on COVID-19.

On the other hand, since January 28, on which Taiwan had the first confirmed local case, the number of public opinions about COVID-19 reached the peak (i.e. 10,722) on news media on January 30. Moreover, the increase in the number of public opinions took on a regular pattern. During the weekdays, the number of Internet public opinions on news media all exceeded 10,000. On weekends, the number was all greater than 6,000. The changes in the number are presented in Figure 1.

Further analysis found that when a relatively large number of Internet public opinions is often attributed to the occurrence of an event, which results in a surge in Internet public opinions on the same day or the next few days. For instance, the highest point for the number of Internet public opinions on COVID-19 on a day appeared on March 19 ( $n = 21,511$ ) because “hoarding chaos” happened on the same day.

Figure 1. Tendencies for internet public opinions on covid-19 (Compiled by the present research)



During the epidemic, “hoarding as soon as possible” went viral on the Internet, which led a massive number of people to rush to purchase daily necessities. On March 19, an Internet user named tombknight asked about the necessity of hoarding on PTT gossiping (Figure 2 and 3). In the post, tombknight mentioned that the elderly asked the family to store a-month-worth food and withdraw deposits from banks. The post did not trigger much discussion, and a majority of the comments showed that they did not believe that it was necessary to do so.

The analysis of keywords can reveal the event that netizens discuss on the Internet or the words that news media use in their reports during the COVID-19 pandemic. The words mentioned the most frequently include the name of the disease, such as Wuhan pneumonia ( $n = 2,555,411$ ) and novel coronavirus ( $n = 1,281,097$ ), the official organization, such as the Central Epidemic Command Center (CECC) ( $n = 2,330,279$ ), and the name of hard-hit countries, like China ( $n = 441,844$ ) and America ( $n = 333,058$ ). The words on word clouds were frequently brought up during the COVID-19 pandemic,

Figure 2. Hoarding-related post on March 19 (Compiled by the present research)

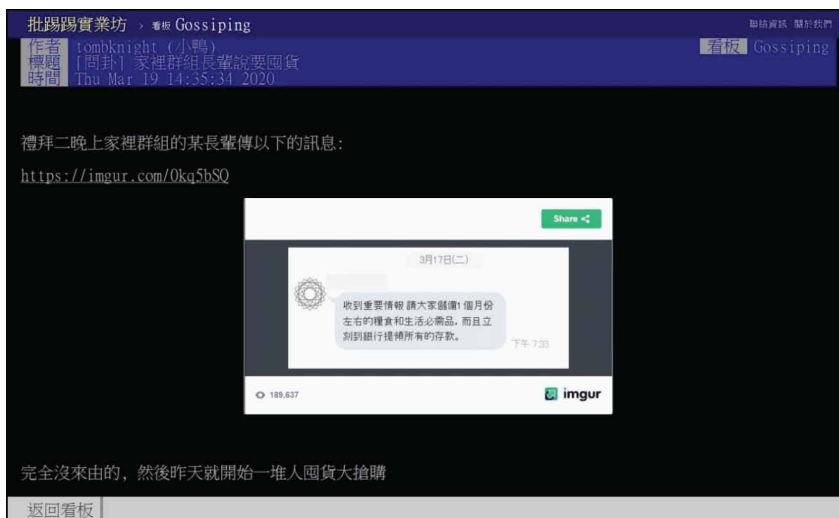
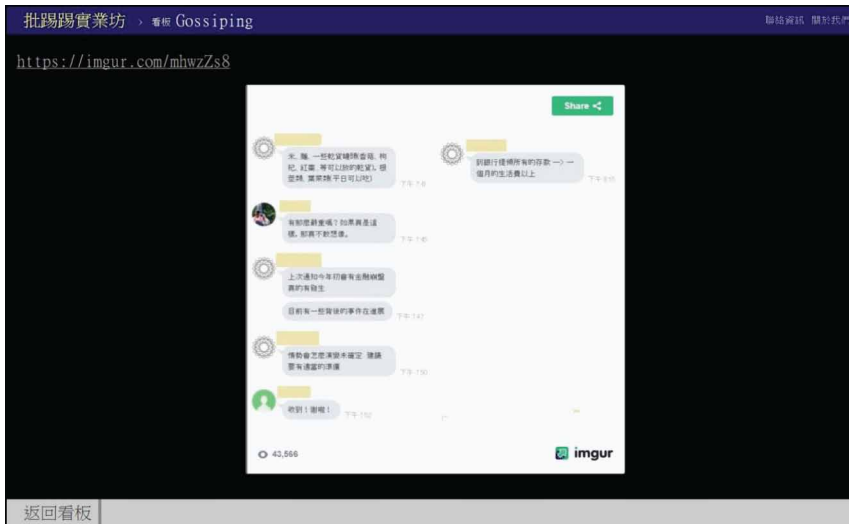


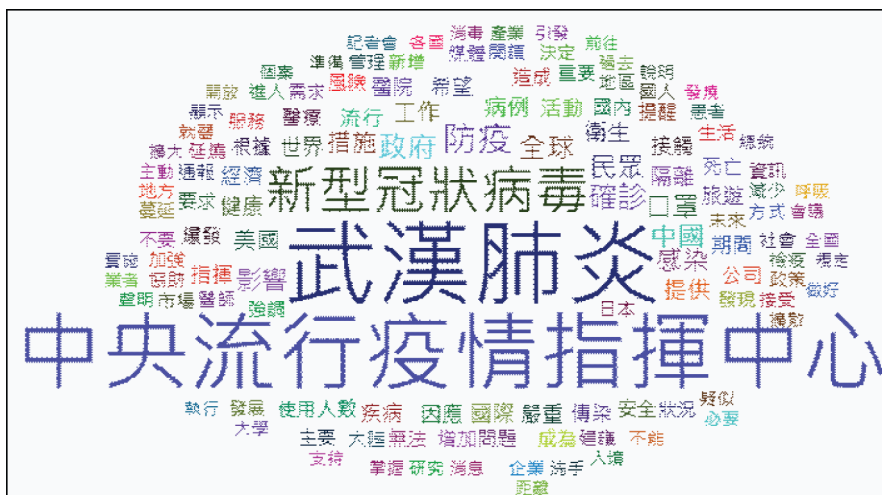
Figure 3. Hoarding-related post on March 19 (Compiled by the present research)



as shown in Figure 4, and the top 20 words that often appear in the Internet public opinions on the COVID-19 pandemic are listed in Table 2.

Emotional analysis of the Internet public opinion on COVID-19 reveals the emotions that netizens' opinions convey, as shown in Figure 5. COVID-19 was called Wuhan Pneumonia in the beginning, which caused netizens' negative emotions to remain intense. With the decrease of confirmed cases in Taiwan, their negative emotions was getting less intense from the peak (84.18%,  $n = 165$ ). On the other hand, officials and soldiers on Dunmu Fleet, a navy warship, were confirmed infected between March 18 and March 23, which caused negative emotions to intensify (greater than 50%). Subsequently, the pandemic was taken under control, and on May 3, no confirmed cases was found for the first time since March 23. On May 4, the lines of the proportions for the words used to convey

Figure 4. Word cloud for the keywords in internet public opinions on Covid-19 (Compiled by the present study)

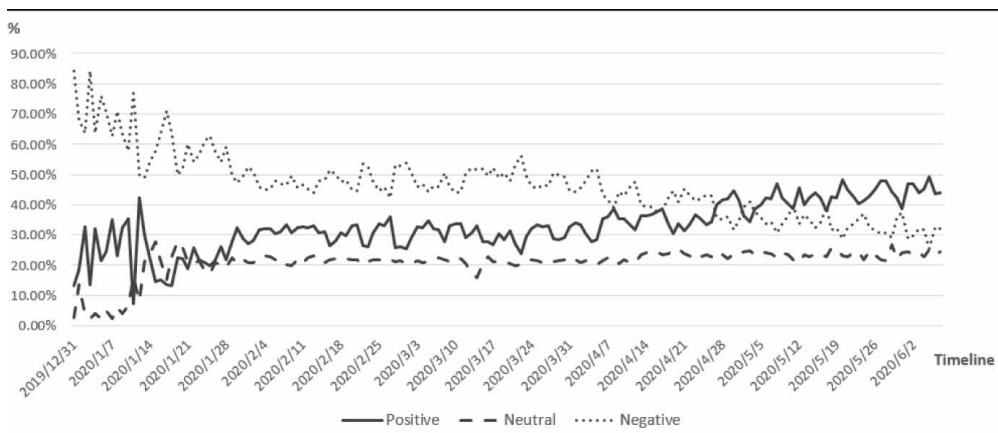


**Table 2. The top 20 keywords in internet public opinions on COVID-19**

Keywords about Internet Public Opinions on COVID-19	Count (n)
Wuhan pneumonia	2,555,411
CECC	2,330,279
Novel coronavirus	1,281,097
Epidemic prevention	667,416
Definitive diagnosis	509,752
Government	470,083
The whole globe	458,540
The public	442,113
China	441,844
Masks	435,606
Influence	401,607
Infection	392,776
Measures	371,398
America	333,058
Work	321,279
Hygiene	317,455
Cases	292,979
Quarantine	279,158
Health	276,610
International	272,476

Source: Collated by the present study

**Figure 5. Emotions in internet public opinions on COVID-19 (Compiled by the present study)**



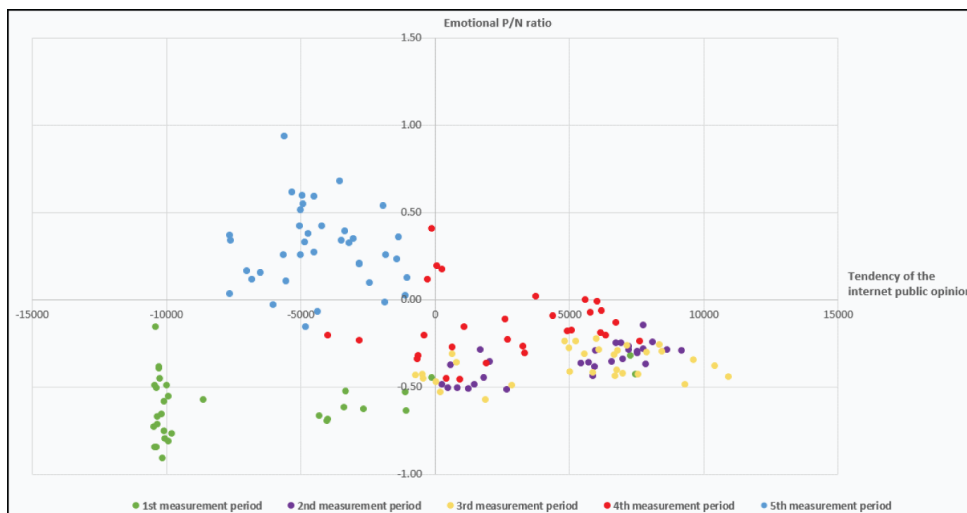


positive (38.69%,  $n = 3,349$ ) and negative (37.75%,  $n = 3,268$ ) emotions crossed; in other words, their percentages equaled. Since that, the positive emotions remained stronger than the negative one until June 7, the finishing time of the present study, so the ratio of P to N remained greater than 1.

Moreover, the emotions in the public opinions on COVID-19 were categorized based on time periods<sup>2</sup>, which are made into perceptual maps<sup>3</sup>, as shown in Figure 6. It can be found that during the first period, the emotional perception fell in the third quadrant, which was small in the number of public opinions and low in the ratio of P/N, for the epidemic was not clear and COVID-19 was named Wuhan pneumonia in this phase. Between the second and fourth periods, the emotional perception mainly fell in the fourth quadrant, which is large in the number of public opinions and low in the ratio of P/N, for the epidemic was becoming worse and the events like “mask chaos” and “hoarding chaos” happened. In the fifth period, the emotional perception fell in the second quadrant, which was small in the number of public opinions and high in the ratio of P/N because the COVID-19 pandemic was taken under control in Taiwan, so the emotions were positive and the discussion about it decreased.

As shown in Table 3, further analysis of the emotions in the posts on news media, social media and discussion boards, can be found that news media were in a relatively neutral position ( $P/N = 0.92$ )

**Figure 6. Emotional perception of the internet public opinions on COVID-19 (Compiled by the present study)**



**Table 3. Emotions in internet public opinions on COVID-19 on different media**

Variables	Positive, n (%)	Neutral, n (%)	Negative, n (%)	P/N ratio
News (n = 867,647)	312,814 (36.05)	216,422 (24.94)	338,411 (39.00)	0.92
Social Media (n = 476,766)	143,085 (30.01)	71,126 (14.92)	262,555 (55.07)	0.54
Forum (n = 79,886)	16,213 (20.30)	24,510 (30.68)	39,163 (49.02)	0.41
Total (n = 1,424,299)	472,112 (33.15)	312,058 (21.91)	640,129 (44.94)	0.74

Source: Compiled by this study

when reporting COVID-19 events. As for social media, although news media repost their reports on their social media account, Internet users conveyed stronger negative emotions on social media, for they are able to freely comment on the post. With regard to the discussion board, netizens often send radical posts or comments, so its P/N was 0.41, indicating that stronger negative emotions were seen on the discussion board.

Furthermore, the analysis of the replies to some comments found that suspected netizens from Chinese mainland participated in the reply. One reply read, “Our Taiwan of China should have faith in the authorities. Everything will be fine.”, as shown in Figure 7 and 8, while the other read, “Thanks, Taiwan Province of China”, as shown in Figure 10 and 11. When these two posts appeared, Taiwanese netizens flocked to comment on the posts.

The identities of the two suspected Chinese netizens were analyzed, and it was found that they might be part of the Chinese cyber army, as shown in Figure 9 and 12. The Twitter accounts of a cyber army have two characteristics: one is that the account has few friends, while the other is that their posts are sent during office hours (Hsin-Chang Jian, & You-Ju Li, 2019). The profile of the above two accounts and the time when their posts were sent conform to the above two characteristics. That is why the present study presumes that they are part of the Chinese cyber army.

The present study conducted correlation analysis, expecting to explore the correlation between different media, as shown in Table 4. The Pearson correlation analysis revealed that the correlation coefficients fell between 0.842 and 0.954, indicating highly positive correlations between the media. In addition, it can be found that when the number of Internet public opinions on COVID-19 increased on a media, the number on the other two showed a similar tendency.

Figure 7. The posts sent by a suspected account of the Chinese cyber army (Compiled by the present research)



Figure 8. The posts sent by a suspected account of the Chinese cyber army (Compiled by the present study)



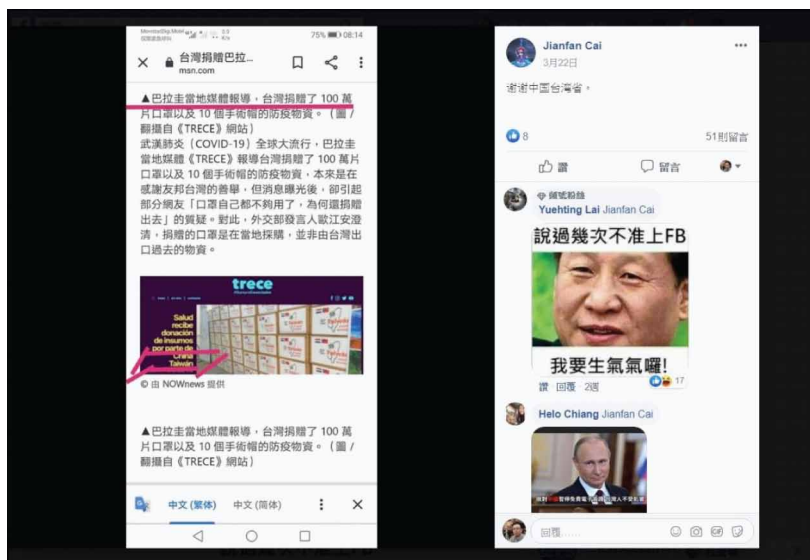
Figure 9. An account suspected to be part of the Chinese cyber army (Compiled by the present research)



Figure 10. The posts from a suspected account of the Chinese cyber army (Compiled by the present research)



Figure 11. Posts sent by a suspected account of the Chinese cyber army (Compiled by the Present Study)



## 5. CONCLUSION

The analysis of Internet public opinions on COVID-19 reveals that the public opinions mainly appear on news media, which is followed by social media like Facebook, Youtube, and Instagram, and forums respectively. During the research period, there were an average of 6,838 reports and discussions about COVID-19 on news media. Since January 30, 2020, on which the first peak of the number of

Figure 12. A suspected account of the Chinese cyber army (Compiled by the present study)



Table 4. Correlation between internet media regarding tendency of public opinions

Media	News	Social Media	Forums
News	-	0.954**	0.842**
Social Media	0.954**	-	0.941**
Forums	0.845**	0.941**	-

\*\* P<0.01

public opinions appeared, the Internet public opinions on COVID-19 presented a regular pattern. On weekdays, the number of public opinions on news media exceeded 10,000, while the number exceeded 6,000 on weekends, for related organizations conducted press conferences and made announcements when the COVID-19 pandemic began, which influenced the reports of news media.

The analysis of the keywords related to the COVID-19 pandemic revealed that the focus words were highly related to COVID-19, either in the reports of news media or on the discussion board, such as Wuhan pneumonia, CECC, novel coronavirus, epidemic prevention, and definitive diagnosis.

Further analysis of the emotion conveyed by keywords revealed that some words had been deemed as ones that convey negative emotions like Wuhan pneumonia, but they appeared in COVID-19 posts that conveyed both positive and negative emotions. Therefore, it can be regarded as a word that convey neutral emotions.

In terms of the emotions of public opinions on COVID-19, focus words used to convey negative emotions accounted for 84.18% because COVID-19 is a worldwide pandemic and the virus was first believed to originate from China. The lines of the percentages for the focus words that convey positive and negative emotions crossed on May 4, 2020, because Taiwan gradually took under control the

COVID-19 pandemic, and the confirmed cases only occurred in a small scale, either for overseas exported cases, local cases or for cluster infections on Dunmu Fleet. Subsequently, the percentage of the focus words that convey positive emotions remained higher than that of negative ones between May 5, 2020 and June 7, 2020, the time when CECC announced “unsealing Taiwan”, and the ratio of P to N remained greater than 1.

Lastly, the peaks of the number of Internet public opinions on COVID-19 were compared with relevant events, showing that the events, such as the occurrence of confirmed cases, rushing to purchase masks, and hoarding chaos, all increased the discussion on COVID-19 online. The correlation analysis revealed that news media, social media and forums were highly positive correlated. Therefore, it can be concluded that COVID-19 events affect the changes in the number of Internet public opinions.

## REFERENCES

- Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2014). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, 79–80, 3–15.
- Di Gennaro, F., Pizzol, D., Marotta, C., Antunes, M., Racalbutto, V., Veronese, N., & Smith, L. (2020). Coronavirus Diseases (COVID-19) Current Status and Future Perspectives: A Narrative Review. *International Journal of Environmental Research and Public Health*, 17(8), 2690. doi:10.3390/ijerph17082690 PMID:32295188
- Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising Qualitative and Quantitative Evidence: A Review of Possible Methods. *Journal of Health Services Research & Policy*, 10(1), 45–53. doi:10.1177/135581960501000110 PMID:15667704
- Du, S. (2020). *The Tendency and Rise and Fall of Taiwanese Consciousness of Unification and Independence: A Comparison of Three Search Engines of the Databases*. Hanlu Book Publishing.
- Grendstad, G. (2007). Causal Complexity and Party Preference. *European Journal of Political Research*, 46(1), 121–149. doi:10.1111/j.1475-6765.2006.00689.x
- Huang, W., Li, R., & Meng, J. (2015). Study on Dissemination Elements and Operational Mechanism of Multimedia Network Public Opinion under the Big Data Environment. *Library and Information Service*, 59(21), 38–44.
- Jian, X., & Li, Y. (2019). *How Twitter Recognizes Cyber Army Accounts?* READr: [https://www.readr.tw/post/2029?fbclid=IwAR1lnz-5V\\_Esl8iRdKPZOLTjN8hGWyLEGt3wE7QLSfLQj2HZH-GSSDc9m50](https://www.readr.tw/post/2029?fbclid=IwAR1lnz-5V_Esl8iRdKPZOLTjN8hGWyLEGt3wE7QLSfLQj2HZH-GSSDc9m50)
- Johns Hopkins University. (2020). *COVID-19 Dashboard*. <https://coronavirus.jhu.edu/map.html>
- Ministry of Health and Welfare. (2020). *COVID-19 Epidemic Prevention Network*. <https://covid19.mohw.gov.tw/ch/mp-205.html>
- Nan, F., Suo, Y., Jia, X., Wu, Y., & Shan, S. (2018). Real-Time Monitoring of Smart Campus and Construction of Weibo Public Opinion Platform. *IEEE Access: Practical Innovations, Open Solutions*, 6, 76502–76515. doi:10.1109/ACCESS.2018.2883799
- Ragin, C. C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. University of California Press.
- Ragin, C. C. (1999a). The distinctiveness of case-oriented research. *Health Services Research*, 34(5), 1137–1151. PMID:10591277
- Ragin, C. C. (1999b). Using qualitative comparative analysis to study causal complexity. *Health Services Research*, 34(5), 1225–1239. PMID:10591281
- Ragin, C. C., Shulman, D., Weinberg, A., & Gran, B. (2003). Complexity, Generality, and Qualitative Comparative Analysis. *Field Methods*, 15(4), 323–340. doi:10.1177/1525822X03257689
- Rihoux, B. (2003). Bridging the Gap between the Qualitative and Quantitative Worlds? A Retrospective and Prospective View on Qualitative Comparative Analysis. *Field Methods*, 15(4), 351–365. doi:10.1177/1525822X03257690
- Rihoux, B. (2006). Qualitative Comparative Analysis (QCA) and Related Systematic Comparative Methods: Recent Advances and Remaining Challenges for Social Science Research. *International Sociology*, 21(5), 679–706. doi:10.1177/0268580906067836
- Schneider, C. Q., & Wagemann, C. (2006). Reducing complexity in Qualitative Comparative Analysis (QCA): Remote and Proximate Factors and the Consolidation of Democracy. *European Journal of Political Research*, 45(5), 751–786. doi:10.1111/j.1475-6765.2006.00635.x
- Sourcing Big Data. (2019). uMiner Manual. Taipei: Sourcing. *Big Data*.
- Zhang, Y. (2009). The Definition and Differentiation of “Public Opinion” and Relevant Concepts. *Zhejiang Academic Journal*, 2009(3), 182–184.


## ENDNOTES

- <sup>1</sup> Netizen refers to the person who initiates and participates in the activities that form public opinions. It was proposed by Michael Hauben, who believes that netizens are a group of Internet users who have a community consciousness and behavioral connections with each other. “Community” here is not defined by the general sense of geographical areas.
- <sup>2</sup> This research divided the periods of Internet public opinions based on times. The first period fell between December 31 and January 31, the second fell in February, the third fell in March, the fourth fell in April, and the fifth fell between May 1 and June 7.
- <sup>3</sup> The X-axis of the perceptual map presents the tendency of the public opinion while the Y-axis stands for the ratio of P to N. The average number of public opinions ( $n = 10,600$ ) and the neutral emotion P/N equals to 1, which was set as the origin of the coordinates, and accordingly, other related values were used to set coordinates and made into perceptual maps.



# Ontology-Based IoT Healthcare Systems (IHS) for Senior Citizens

Sakshi Gupta, Birla Institute of Technology, Mesra, India

 <https://orcid.org/0000-0001-7793-9858>

Umang Singh, Institute of Technology and Science, Ghaziabad, India

## ABSTRACT

Rapid incremental growth in population increases the virulence of infectious diseases worldwide. Due to this, health hazards with population growth raise pollution in the air, water, and soil and affect the immunity of individuals. To handle the situation, reliable and easy-to-reach healthcare services are required. The proliferation of connected technologies along with the internet of things (IoT) are providing modern healthcare with extensive care. All-pervading IoT technology is gaining attraction nowadays. This paper presents a brief about the e-healthcare system along with its framework. This attempt also presents the ontology approach as data produced by healthcare applications is vast and unstructured and needs to be organized into a proper format with a smooth flow of data to result in less request-response time. Further, this paper discusses the impact of disease on senior citizens in the current scenario.

## KEYWORDS

Dataset, Diseases, Elderly People, Healthcare System, Infections, IoT, Ontology, OWL, Risk Factors, URI

## INTRODUCTION

A reliable and prompt E-healthcare system is the immense requirement of today's scenario. The world is facing a pandemic COVID-19 (Coronavirus Disease-19) (Recalcati, 2020) situation and exploring the possibilities to find out the related vaccine as soon as possible. As this disease has symptoms like fever, fatigue, cough, anosmia, ageusia, and human to human transfer, though no standard vaccine has evolved to date. Studies show that senior citizens are at high-risks of attacked by COVID-19 as their immune system is low. It is difficult for them to go to hospitals and labs for a routine health check-up or in an emergency. An E-healthcare system (electronic healthcare system) (Maglogiannis, 2009) requires that we can collect all the data of patients, doctors, nurses, and other actors directly related to healthcare. Data can be gathered from hospitals, labs, and sometimes from sensors attached to patient's bodies or nearby bodies. At the data acquisition layer, data should be acquired from a meaningful perspective so that it can be re-used without less filtration process. Analysis of that data is necessary to find out the existence of any disease. Various E-healthcare record systems are overviewed in (Yadav et al., 2018). The Internet of Things (IoT) has already endowed its application

DOI: 10.4018/IJBDAH.287604

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

as a smart healthcare system or intelligent healthcare system sometimes termed as IoMT (Internet of Medical Things). IoMT is a collaboration of medical things and applications to provide connectivity to healthcare systems over the internet. IoT helps to access healthcare data quickly for large scale applications. This paper is organized as follows. Section II provides some related works. Section III presents IoT healthcare systems and Risk factors for elderly people. Section IV describes the framework of IoT healthcare system and related methodology. Section V presents research findings from the current scenario related to Covid19. The final section concludes our study.

## LITERATURE REVIEW

Tun et. al (2020) discussed an overview about 11 different applications of IoT and wearables like aged care monitoring, chronic patient healthcare monitoring, the clinical applications, emergency conditions, mental health, and others to supervise healthcare of senior citizens and pointed out clinical point of view in comparison to technology point of view. Authors Lee et al., (2020) focused on older people and people with disabilities, caregivers, and healthcare providers with the help of a face-to-face questionnaire. The authors mentioned that discrepancies are still present in the current scenario related to the requirement of IoT healthcare services. As people require these technical services mostly in emergencies as well as persons with mild disabilities. In Pinto et al., (2017) a complete IoT healthcare system has been provided for monitoring the health life of elder persons and trigger alarm in case of any emergency. A wrist band is used by an elder person can collect data and send it to the We-care server with the help of a 6LoWPAN protocol. An Elderly IoT healthcare system is proposed in Park et al., (2017) where authors focus on brain stroke issues. Parameters like blood pressure, pulse rate, sugar level, oxygen level, motion tracking are tracked in senior citizens via wearable devices to control any emergency. Ray et. al (2014) proposed a five-layered framework named home health hub IoT (H3IoT) for elderly people who are in a homely environment. Basanta et. al(2016) presented solutions for monitoring the health of elder citizens with the integration of IoT technology. Authors have presented their health issues in older natives related to physical disability and severe psychological depressions.

Related work mentioned in [14-23] is that the organization of congregate data is one of the focal challenge as most data is available in a heterogeneous format. A semantic and flexible data model is required to give a smooth flow in data. For this, ontology is used to explain the properties of a domain by describing concepts and relationships among entities of the domain. Entities in IoT like IoT devices, IoT applications, users, developers are not able to use homogeneous platforms all the time, so ontologies provide a consistent and formal representation of data among these entities. Here, semantic web comes into the picture. According to the World Wide Web (W3C) - the “*semantic web provides a common framework that allows data to be shared and reuse across applications enterprise and community*” [24]. Semantic web technologies like- Resource description framework (RDF), Web Ontology Language (OWL) are used to link data over the web in such a way so that machines can understand and manipulate data. The semantic web in the IoT field sometimes referred as the semantic web of things (SWoT) is responsible for representing and describing the things on the web via Uniform Resource Identifier (URI) or Internationalized Resource Identifier (IRI) and relationship is defined among these things by RDF [25]. Ontology provides domain knowledge and standardization as well as codification of that knowledge in a machine-understandable format so it can be reused by people, databases, and applications [26]. According to the W3C consortium, the Semantic web can be used to integrate varied data in one seamless application, as it provides a common framework that allows the sharing and reusing of data on the web. Ontology opens new ways of data integration and analysis in the IoT healthcare system where diverse data can be expressed and amalgamated by using some ontology languages like ontology web language (OWL) and resource description framework (RDF).

In [27] author discussed a semantic middleware data model to integrate IoT healthcare information systems to electronic healthcare records (EHR) with the help of ontology. Two components, semantic

**Table 1. Comparison of technologies and methodologies identified as a solution for diseases in the existing scenario**

S.No	Existing Proposed Model	Identified diseases	Methodology Used	Process Flow	Data collection methods	Factors still unexplored
1.	The IoT-based heart disease monitoring system for pervasive healthcare service. (2017) (Li et al., 2017)	Heart diseases (blood pressure, ECG, heart rate, pulse rate, blood glucose, SpO2, and blood fat)	Based on data acquisition and data transmission.	A Prototype can monitor the patient's physical parameters and transmit them. The Smartphone is used as a connector and a web-based application is used for doctors to monitor data.	Parameters and frequency of parameters are designed after interviewing medical experts. Physical signals are collected via various sensors. ECG-128Hz (Sampling frequency) BP-2seconds SpO2-2sec	Not considered for the elderly age group
2.	Internet of things–based personal device for diabetes therapy management in ambient assisted living (AAL). (2011)(Jara et al., 2011)	Diabetes mellitus in elderly people	Based on therapy model including technologies- RFID, 6LoWPAN	Self-monitoring operation is referred for blood glucose. Information is received through instructions by doctors and nurses at home.	Diabetes Information System from internet	Not defined any security
3.	An E-health system for monitoring elderly health based on Internet of Things and Fog computing (Hassen et al., 2019)	Physiological and general parameters like – ECG, Body temperature, BP, oxygen level, pulses are collected	Proposed an e-healthcare system for elder people. Based on IoT and Fog computing.	Data is analyzed as well as stored and send to the cloud via REST API. The resulting feedback is found high after providing awareness users.	Used the HW V2 platform and an android app for fog computing.	Not focused on any specific disease
4.	Identifying risky environments for COPD patients using smartphones and the internet of things objects. (2014) (Kouris & Koutsouris, 2014)	Chronic obstructive pulmonary disease (COPD)	Based on the data processing scheme and cloud-hosted services.	Successfully identify risky environments on a small scale. Sensors are deployed around patients so that data can be received through a smartphone.	MATLAB, WEKA, wearable devices using Arduino	Not focused on elderly
5.	Smart Portable Monitoring Device for Asthma Patients (2016) (Abinayaa & Raja, 2016)	Asthma disease	The proposed system is used to track real-time data of asthma patients and instructions are delivered by medical staff accordingly. Based on android application software and a Wi-Fi module.	Real-time data of asthma disease like- body temperature, Air pressure, Carbon Dioxide are tracked and get instructions by doctors.	Temperature sensor, accelerometer, Rpi 2 card. Real-time data collection of patients.	Not focused on elderly
6.	Remote health monitoring of the elderly through wearable sensors (2019) (Al-Khafaji et al., 2019)	Physiological and general parameters are collected for senior citizen in Ambient Assistive Living (AAL) environment Parameters- oxygen saturation, pulses, BP	Proposed an (SW-SHMS) smart health home-based monitor system to gather real-time data. Based on smart phone, technologies- ZigBee, Bluetooth, and wearable devices like- oximeter, smart watch	Faster response time, fewer packet drops, and low Real-time data gathered in patient's layer, data send to cloud layer and Doctors will monitor patients' data via hospital layer	Wearables, smartphone, heartbeat sensor, Arduino UNO device Real time data collection of patients via Wearables, smartphone, heartbeat sensor, Arduino UNO device	Not any specific diseases
7.	A diagnostic prediction model for chronic kidney disease in the internet of things platform (2020) (Hosseinzadeh et al., 2020)	Chronic Kidney Disease (CKD)	Data is collected using the IoT platform and then data mining algorithms is applied to predict CKD. Based on IoT platform and data mining algorithms SVM, Naïve Bayes	The decision tree algorithm gives the best result in terms of accuracy, sensitivity, and specificity in applying the dataset. Feature selection approach is applied on dataset for CDK prediction	Dataset 1- <a href="https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease">https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease</a>	Not focused on elderly

*continued on following page*

**Table 1. Continued**

S.No	Existing Proposed Model	Identified diseases	Methodology Used	Process Flow	Data collection methods	Factors still unexplored
8.	Adeniyi Onasanya, Maher Elshakankiri [Smart integrated IoT healthcare system for cancer care (2019)] (Onasanya & Elshakankiri, 2019)	Cancer disease	Proposed IoT based healthcare framework is discussed for cancer patients, emphasis on cancer care services and business analytics and cloud services Based on IoT healthcare framework and cloud services	Framework solely for cancer care patients with treatment options, security aspects covered. Smart IoT-Enabled healthcare system is proposed with sensor layer, data layer, hospital layer and cancer layer.	IoT devices- sensors, actuators, WSN, IoT technologies- NFC, BLE, LPWA, ZigBee. Real time data collection.	Not focused on elderly
9.	Johannes Mae, Endra Oeyand Ferdian Stanley Kristiady[ IoT based body weight tracking system for obese adults in Indonesia using real-time database] 2020 (Mae et al., 2020)	Obesity	A system " Digital weight scale" is implemented to monitor weight time to time with the use of IoT. Based on IoT and smart phone application is modified by digital weight scale.	The weight scale gives 99% accuracy with real-time data and 40 hours of usage Weight scale is used to measure weight and data upload to cloud and access using application.	A weight scale unit, smartphone applications, and real-time database Real-time database	Not focused on elderly
10.	Shah, S. T. U., Badshah, F., Dad, F., Amin, N., & Jan, M. A. (2019) [Cloud-Assisted IoT-Based Smart Respiratory Monitoring System for Asthma Patients] (Shah et al., 2019)	Asthma	proposed a cloud-assisted IoT-driven healthcare monitoring framework for asthma patients. Based on IoT services and cloud-assisted healthcare system and feature extraction process	Remote monitoring of patients health and data analysis at cloud.	Simulation is done via VM as java based simulator, a virtual machine with Intel Core i3 1.7 GHz processor, 3GB DDR ECC RAM, 8MBPS bandwidth, and running windows server 2012. Dataset- <a href="https://www.physionet.org/content/bidmc/1.0.0/">https://www.physionet.org/content/bidmc/1.0.0/</a>	Not focused on elderly

EHR triple store, and semantic IoT triple store are coupled in traditional semantic architecture where reasoning and querying are performed in EHR by OWL and some existing ontologies like SSN ontologies, Geo ontologies are used to represent sensor resources and data in IoT triple store. [28] proposed Wearable Healthcare Ontology (WH\_Ontology) that aggregates data generated from wearable devices to create valuable knowledge and take the appropriate decision.

Though lots of literature work is exist for integration of semantic technologies and ontology in IoT healthcare systems but very limited studies are available for integration of ontology in IoT healthcare applications for senior citizens. In today's scenario, this is a very prominent challenge that how data should be presented about aged people over the web, so that they can get quick and easy health care services to improve their lifestyle and physical as well as mental health. [29] presents a system named vINCI that provides personalized healthcare services of older adults by exploiting patient's profile using ontologies which ensure good quality of life and automated monitoring for older people but lacking an approach of pre-processing of the initially stored data at an early stage, so that more clear ontologies can be developed. In current worldwide scenario, provision of vaccination is responsible for improved health and prevented 6 million death annually. Impact of vaccines in SARSCoV2 depends on three aspects (1) Health: Eradicate infectious diseases, Immunity; Cancer prevention, vaccine preventable diseases (2) Economic: Cost of Vaccine, Awareness Programme (3) Social: Strengthen healthcare and Empowerment of women.

This article proposed the architecture of an ontology-based IoT healthcare system for senior citizens where pre-processing of data is done at the network edge.

## IOT HEALTHCARE SYSTEM (IHS)

IoT is one of the emerging technologies that can be considered as a network of interconnected devices to share data with each other without the intervention of any human. It is believed that by 2025 more than 21 billion devices will be connected in the entire world. The healthcare sector is one of the biggest sectors in IoT where devices are directly connected to the human body. IHS offers convenient healthcare services to patients and doctors at their places to diagnose a patient's health and contains the information of patients, doctors, nurses, types of diseases. IHS works on layered architecture, consisting of data perception or data acquisition layer as a base layer where sensors are connected to the human body or nearby humans like smartwatches, smartphones, fitness trackers, smart clothing, and implantable [30].

These miniatures sense and track heartbeat, oxygen saturation level, body temperature, heart rate, pulses, and other activities going on in the body. Data collected at this layer will forward to mobile devices or directly to the cloud from where doctors can access that patient's data and diagnose the health issues (Fig.1(a)). Cloud is a group of data centers, hosts, VMs, resources. Data centers hold various resources and lists of different applications to store the data. Integrating cloud in IoT healthcare opens new ways for rising technologies like Machine Learning, Artificial Intelligence, and data mining to analyze the data and drag valuable results from patient's real-time data so that a quick and better decision can be taken at the right time. Figure 1(a)-(b) depicts a very basic healthcare system where information produced by sensors transmits to server and server then sends all data to the cloud. If processing is required, all data processing is done in the cloud and health care experts can retrieve the patient's data from the cloud and give their feedback directly to patients.

Figure 1a. IoT Healthcare system

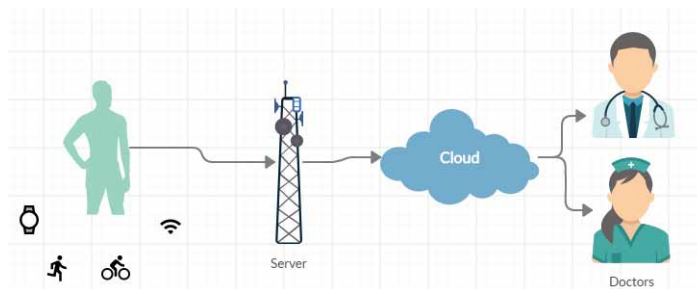
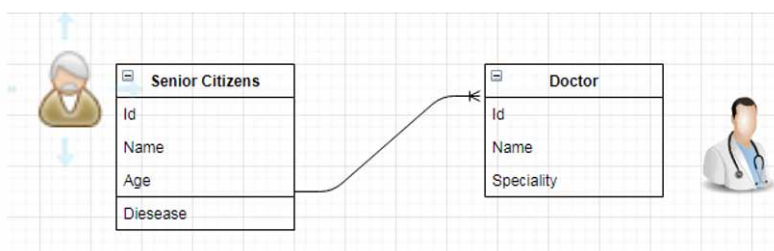


Figure 1b. Class representation in IoT healthcare



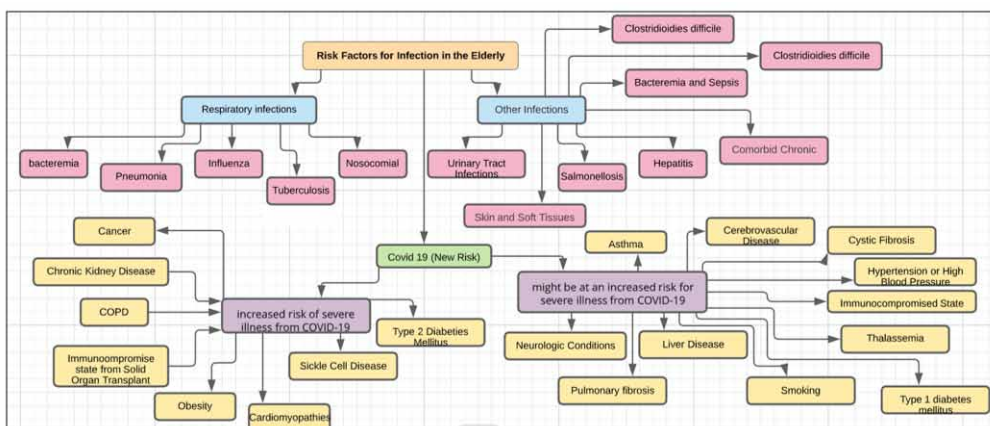
## HIGH RISK FACTORS FOR ELDERLY PEOPLE

Over the past few years, some foremost information communication technologies like- IoT, Big Data, Wireless Sensor Network (WSN), Cloud Computing, and Machine Learning (ML) have spread their roots in the sector of healthcare systems. Data generated by healthcare systems are available in varieties at very rapid speed and vast in nature. And concrete and useful information have to be digging out from this cosmic sea of data. These aforementioned qualities of data satisfy all the V's like- volume, variety, viability, value, velocity, the validity of big data [4] so, healthcare data can be considered as Big data. Due to the gigantic nature of information, handling data in the healthcare system possess some challenges as large scale of data, processing of data, cost, interoperability, integration, security, and privacy [5,6]. Fig.1 presents the risk factor of diseases for infection in Elderly people with age group G1(65-74), G2(75-85), G3(>85). Risk factors for infection in the elderly people as mentioned in fig.1, are raised mainly due to infection in respiratory which causes diseases like bacteria, pneumonia, Influenza, Tuberculosis (chest, spine), and Nosocomial. Other infections like UTI (Urinary Tract Infections), Hepatitis(B-), Clostridioides difficile, etc. which may deteriorate the health of elderly people and can badly affect organs for the proper functioning of the body. In current COVID 19 scenarios, the risk is very high for patients who are already suffering from cancer, kidney disease, COPD, obesity, diabetes, and cardiomyopathies. However, Coronavirus can badly impact the patients who are suffering from asthma, liver diseases, thalassemia, cystic fibrosis, and hypertension (high blood pressure).

As data Analysis and integration is herculean task so ontology-based services play an important role to collect, integrate, and interpret data to make better decisions. Healthcare data can be present and accessed in ontology via OWL (Web ontology language). Ontology provides a conceptual and categorical view of knowledge that is extracted from healthcare systems [7]. National Center for Health Statistics (NCHS) has categorized older adults in 3 groups as group 1: 66-74 years “Young-old”, group 2: 75-84 years “Old-old”, group 3- 85 above years “The oldest-old”. In today's scenario, the importance of the ontology approach as data produced by healthcare applications is vast and unstructured which needs to be organized in proper format with a smooth flow of data and also results in less request-response time. Ontology(O) presents a formal and explicit representation of terminology and installation. It is defined as:

$$O = TBOX + ABOX \quad (1)$$

Figure 2. Highly risk factors for Elderly People



where TBOX represents the Terminological Component, which is a conceptualization associated with a set of facts and ABOX presents the Assertion Component which contains assertions on instances.

In this equation (1), TBOX comprises eight tuples ( $C, \leq C, R, \sigma R, \leq R, A, \sigma A, T$ ) which represents disjoint sets of  $C$ (concepts),  $R$ (Relations),  $A$ (Attributes), and  $T$  (Data Types). Further,  $\leq C$  represents concept hierarchy,  $\leq R$  relations hierarchy,  $\sigma R$  represents relations signature which focuses on what concepts are involved in one specific relation of set  $R$  and represents attribute signature which takes the value of certain data type  $T$ . ABOX is an ontology installation in the form of Role and Instances.

These older adults face a reduction in their physical ability and having diseases like high blood pressure, blood sugar, asthma, chances to have brain stroke. This is the main challenge to take care of the well-being of older and to offer them intelligent healthcare systems by which they can monitor their daily routine and get instruction by healthcare providers at the accurate time to make better decisions. The objective and contribution of this article propose ontology for the wellbeing of senior citizens and pre-processing of heterogeneous data at an early stage that is edge network.

## FRAMEWORK OF IOT HEALTHCARE SYSTEM

As demand increases for online healthcare services, it is necessary to develop a system with low cost, more reliability, and security. For this, a firm and smooth framework are required so that data transmission and reception at every layer should take place with ease. Various literature exists with different types of frameworks of HIS. In [31] the author proposed a cloud-IoT based healthcare framework where doctors, patients, and stakeholders can use this framework for their improved results and improved health, respectively. Here, the cloud server provides PaaS and IaaS services to host applications. Authors in [32] proposed a framework for real-time data with a multilayer framework with the incorporation of layer between the sensing layer and cloud so that basic pre-processing can be done at this new layer. [33] presents an IoT based computational framework to monitor the health of patients in mobile environments also and validate this proposed framework with a case study of footballers' heart rate data. Another framework is also proposed for ECG monitoring system where ECG data is collected than signal enhancing and watermarking for security is done during data collection and some machine learning process like features extraction and classification is done at cloud [34]. In figure 4, a basic framework is presented where the base layer is an IHS sensor layer

Figure 3. Ontology Representation

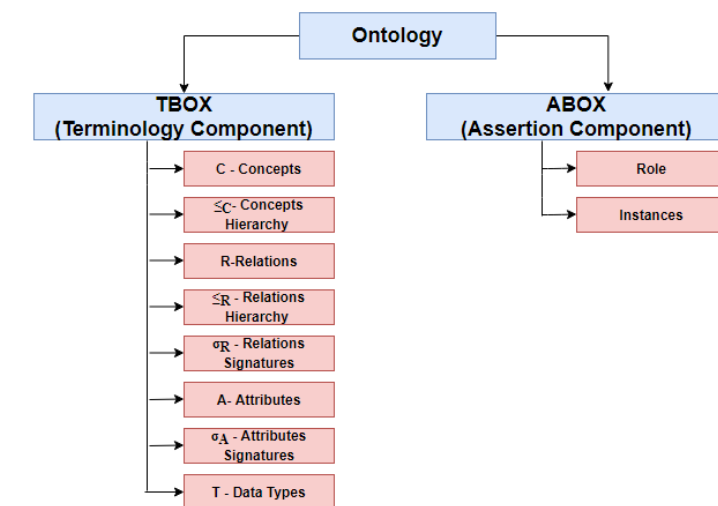
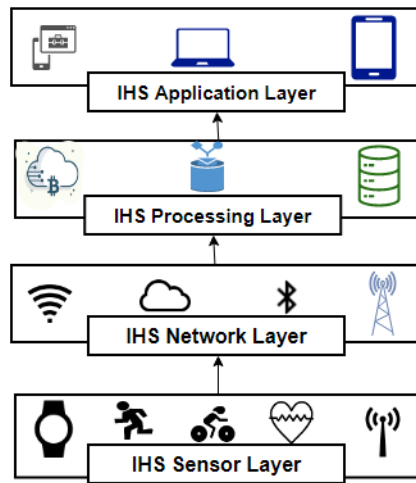




Figure 4. IHS architecture



where deployed sensors create networks among them to sense the body area and acquire data from RFID tags, fit bands, heart rate sensors, and many more.

IHS framework encompasses four layers where the first layer is a sensor layer that incorporated all IoT gadgets like sensors, wearables, smartwatches, actuators, RFID tags. These gadgets are deployed around the physical environment of patients to monitor physiological and vital parameters and data collected through these sensors are sent to the network layer via various IoT technologies like ZigBee, Bluetooth, LPWAN, WI-FI, 6LowPAN. The network layer comprises gateways and edge routers that play the role of a bridge between the sensor layer and the data processing layer. The data processing layer is responsible for applying analytics to raw data gathered by the sensor layer where analytics can be done locally or globally. Local data analysis can be performed at edge router nearer to the devices to divide the burden of the network. Global data analysis can be done in the cloud by applying some data mining and machine learning algorithms to detect the pattern and predict the next values. This analyzed data or useful knowledge will pass to the application layer via the IoT application layer protocol like MQTT, COAP. These protocols present data to users in terms of any desktop applications or any mobile applications. patients can get a view about their signs and get instructions from doctors and nurses.

### Methodology Used for Strengthening Healthcare System

Due to the rapid proliferation of diseases, numerous encoding concepts and ontologies related terms are used in the medical domain. Ontologies like SNOMED (Systematized Nomenclature of Medicine), ICD (International Classification of diseases) offer systematic, computer processable collection of medical terms in humans as well as veterinary medicines to provide encoding terminologies. Further, SNOMED-CT extends support internationally and provides access to terminology and health data related resources in COVID 19 as well [35]. Patients already suffering from diseases Pneumonia, diabetes, asthma, hypertension, cardiovascular, Obesity, chronic kidney have high risk factor of infection from Covid 19 (Table 2).

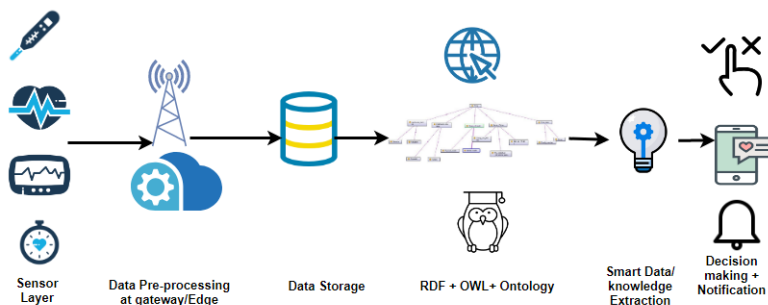
Figure 5 portrays the flow of data, from the generation of data to become smart data for elderly people IoT healthcare applications. This process involves various sub processes that are as follows:



Table 2. Referred Algorithm used for diseases

S. No.	Diseases	Algorithm Used
1.	Glaucoma Diagnosis	Logical Regression [36]
2.	Alzheimer's Disease	Linear Regression [37]
3.	Diagnose Bacterial Sepsis	Random Forest(RF) and ANN[38]
4.	Covid 19	Naïve Bayes Algorithm [ 39] Logistic Regression Algorithm [39] Artificial Neural Network [40] K- Nearest Neighbor [41]

Figure 5. Ontology-based IoT healthcare system for senior citizens



1. **Sensor Layer:** Various sensors are attached to the older adult's body to sense the parameters of their body. These sensors include an oximeter, electrocardiogram, thermometer, fluid level sensor, blood pressure sensor to read the current situation of patients. These sensors have low battery and low memory so used only to sense and transfer the data to the upper layer. No processing is done at this stage. In live tracking of records of individual patient is not satisfactory; then superintending of patient process is executed virtually. For tracing and clustering Data Processing module is initiated.
2. **Data Preprocessing at Edge:** Data generated by sensors are sent to gateways at the network layer through various technologies like Bluetooth, Wi-Fi, ZigBee, LoRaWAN, 5G, LTE. Gateway will store that data packets and sometimes these gateways are available near devices called edges of the network. Basic data processing can be done locally at edge routers. Pre-processing of data involves cleaning, eliminating duplicate packets, normalization of data, basic data aggregation (SUM, AVG, MAX, MIN), and pre-processed data transmitted to the cloud; result in no more burden there. Portable devices like smartphones, smartwatches, compact embedded systems, and compact gateways perform the aforementioned pre-processing tasks.
3. **Data Storage:** This stage can be considered as a warehouse or cloud. Data stored at this stage is huge. Machine Learning, deep learning, and data mining algorithms are applied here to analyze as well as extract knowledge from data.
4. **RDF/OWL/Ontology Layer:** Data stored in the storage is used by the semantic web. Semantic web sanctifies knowledge from data in such a way that improves decision making. Semantic web technologies include RDF and OWL. RDF linked the opened data over the web and present data in triple format "subject-predicate-object" which provides structure and unique identifiers. Ontology presents a knowledge model that identifies a set of concepts also called classes and

tries to define relationships among concepts in a particular domain. The knowledge presented by ontology can be shared and reused at any time. Figure 4 shows the proposed ontology for senior citizens IoT healthcare systems developed by protégé software. Here five classes are identified name- senior citizens, healthcare providers, caretakers, healthcare devices, and types of health. The senior citizen class is further categorized into 3 subclasses as “Young old”, “Old-old”, and “The oldest old”. Similarly, the type of health class is divided into physical health and sub health and other classes are also having subclasses. Properties are defined for different users like senior citizens along with properties name, Id, age whereas doctor has characteristics like id, name, area of specialization. Relationships among classes show links between pairs of classes. For example:

- a. Physician is-a-subclass-of healthcare provider (superclass to subclass relation);
  - b. Senior citizens has-a mental health or physical health (class to class relation);
  - c. Senior citizen uses healthcare devices (class to class relation).
5. **Smart Data:** RDF and OWL are used together to provide smart data, data that can be read and understandable by machine without any human intervention, and able to apply reasoning. When every device and file used in an IoT healthcare system is represented by ontology and linked together over the web then the machine can co-relate and visualize the flow of data, this results in better performance of the system. This data is used for ambulance service; if patient requires hospitalization then ambulance service will be provided and patient information will also be shared with hospital.
6. **Decision Making:** Ontological Meta-data helps healthcare users and healthcare providers to take all necessary steps towards their quality of life and it helps the decision support system (DSS) to decide where to send a notification to patients or caretakers based on ontology-driven advice.

## Research Findings and Comparative Analysis of Worldwide Disease Infections Effect on Elderly Senior Citizens

The coronavirus disease was declared a pandemic by WHO (World Health Organization) on March 13th. The attempt [42] discussed about virus infection and its impact in current scenario as well as recent development. Everyday constantly COVID Cases are rapidly increasing worldwide as time passes while healthcare providers and government bodies are putting efforts to stop the spread of this virus. Europe [43] is one of the most affected continents in which countries like Italy, Spain, the UK, and France. In [44] authors mentioned that the tropical regions are relatively less prone to COVID-19 cases than the European & American regions and transmission of spread is not uniform globally. They also conclude that the death rate is mostly subjective to age group and medical history. Table 3 presents a reasonable analysis of new COVID-19 cases reported in the last 24 hours concerning deaths reported in the last 24 hours on 9th September 2020(WHO). The table 3 presents data from four foremost countries in Covid-19 cases.

**Table 3. Comparative Analysis of Cumulative Cases Vs Deaths reported in 24 hours on September 9th, 2020**

Country Name	Cases Cumulative total	Cases newly reported in the last 24 years	Death Total Cumulative	Deaths newly reported in the last 24 hours	Transmission Classification
India	4,280,422	75,809	72,775	1,133	Cluster of Cases
US	6,222,974	33,486	188,003	462	Community Transmission
Brazil	4,137,521	14,521	126,650	447	Community Transmission
Russia	1,035,789	5,099	17,993	122	Cluster of Cases

The figure 6 presents that the U.S has the highest cumulative cases and India has become the first country in the world with reporting the highest cases and deaths in the last 24 hours. According to the world population prospects, current population of India stood at 1,352,642,280. By 2030, it is expected its population by 1.5 billion people. 35% percent of its population below the age of 25 and more than 65% age below the age of 35. Table 4 presents high level of infections spread among old age people across the globe during Covid 19.

Figure 7(a-c) depicts the worldwide count of COVID 19 cases raised frequently shown in the figure that the Brazil, China, India, USA has the highest cases surpassed confirmed cases for elderly people.

The dataset results mentioned in Fig 9 presents positive cases confirmed per age group. Age groups are binned into the eight groups <G1: under 18, G2: 18-30, G3:31-40, G4:41-50, G5:51-60, G6:61-70, G7: 71-80 and G8: 81+>. Information of cases are classified based on cluster case and community transmission. As per HT reports, most of deaths due to Covid 19 across nation were in age group between 50-70 out of which highest age group was in between 61-70 age among both genders. However, male death ratio is almost 69% and higher as compared to females.

To prevent world from COVID-19 pandemic, doctors, researchers, and government officials are in much action and succeed Within less than 12 months after the beginning of the COVID-19 in developing and authorize vaccine. Vaccines are always helps to save life of million people each year. Vaccine makes human body's immune system strong and prevents from illness. As this pandemic is very fearful and people are dying, so it is immense requirement to develop vaccine against this virus.

After so much of hard work doctors are succeed in developing vaccines and as of 18 February 2021, at least 7 different vaccines have been approved and rolled out in countries where as more than 200 vaccines are in under development [WHO]. There are companies which provide vaccines are Pfizer, Serum Institute of India, Moderna, BioNtech, Fosun Pharma and many more [45]. Vaccination has been started in December 2020 in countries like UK, Russia, China, Israel[Fig.9]. In a study it is find that, one third (33 percent) of people in group who have already coronavirus reported "mild

Figure 6. Situation of COVID 19 Effect in Top 4 most affected Countries (As per report 9th September 2020)

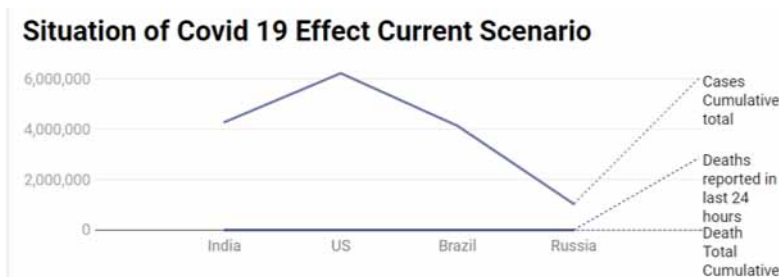


Table 4. High risk infection rate during Covid Scenario in elderly people

Disease Name	Age Group	Covid 19	Affected tract	Organ specific
Bacterial pneumonia	60-65	SARS Cov-2	Respiratory tract	Lung
Elderly influenza	80+	SARS Cov-2, H1N1	Lower and upper both Respiratory tract	Heart, Brain and muscles
Elderly skin infections	65+	HSV type1	Methicillin-resistant Staphylococcus aureus	Internal organs
Gastrointestinal infections	80+	Rotavirus, Adenovirus, Astrovirus, Calicivirus	Gastrointestinal tract	Stomach and Small Intestine

Figure 7a. Mapping of Cases raised frequently country wise Elderly people age (75-84) (Source: <https://covid19.who.int/table?tableChartType=heat>)

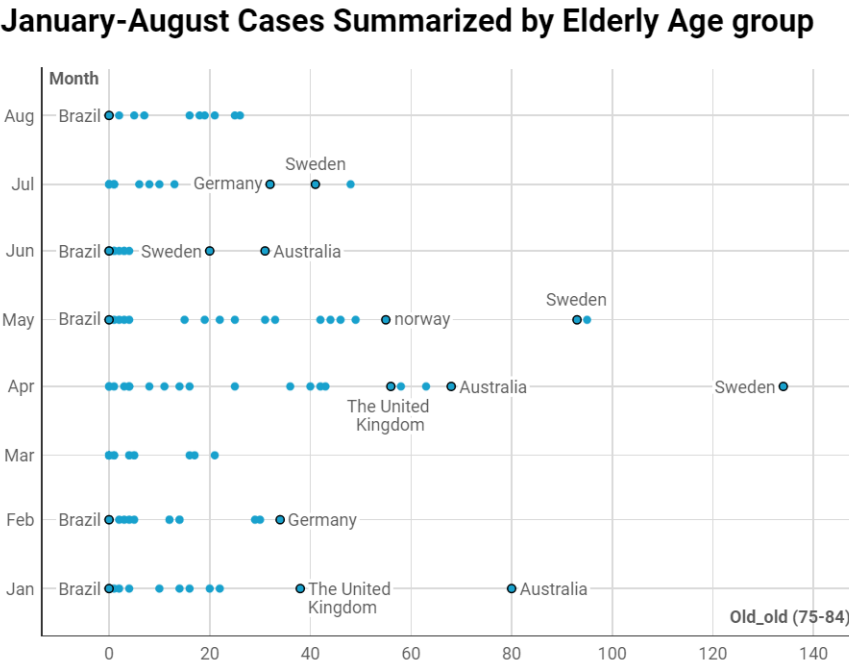


Figure 7b. Mapping of Cases raised frequently country wise Elderly people age (65-74) (Source: <https://covid19.who.int/table?tableChartType=heat>)

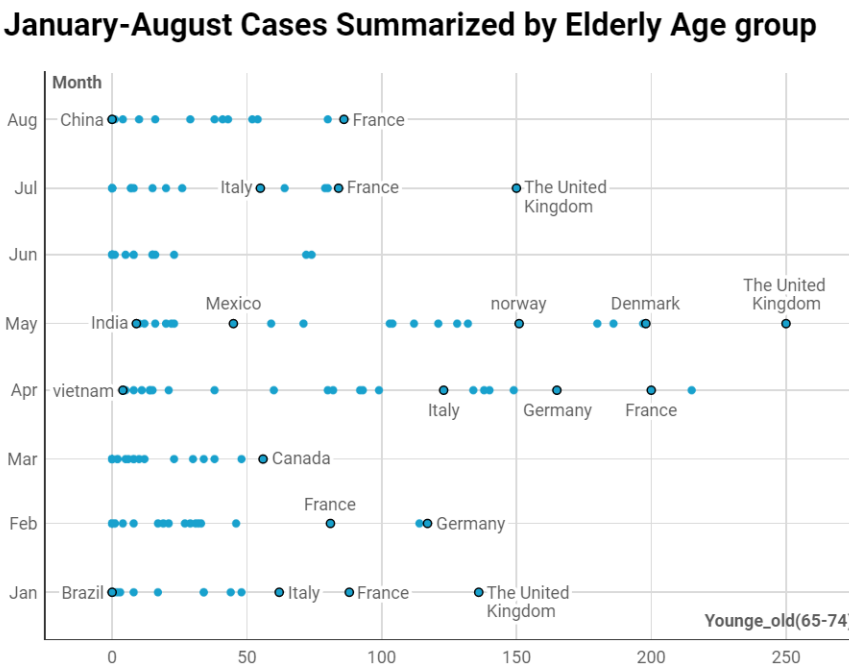


Figure 7c. Mapping of Cases raised frequently country wise Elderly people age (85 Above) (Source: <https://covid19.who.int/table?tableChartType=heat>)

### January-August Cases Summarized by Elderly Age group

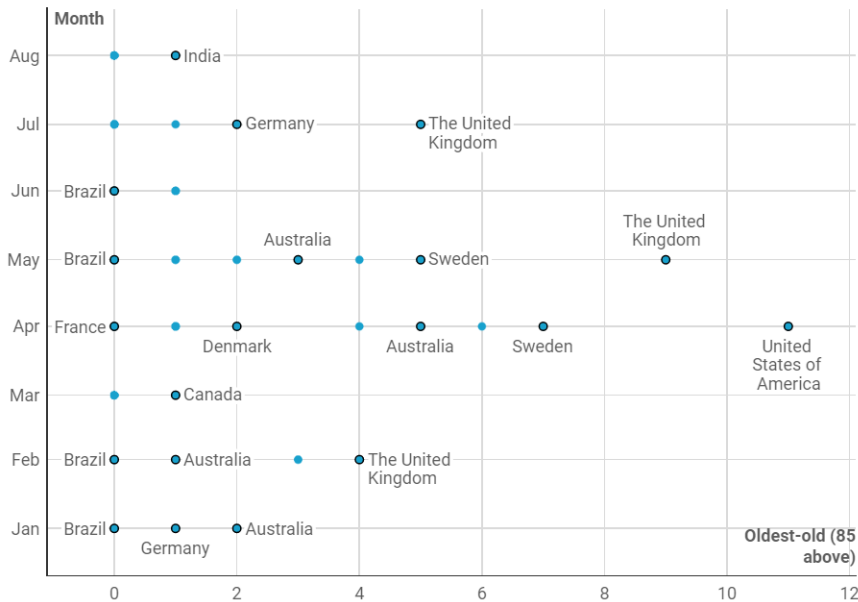
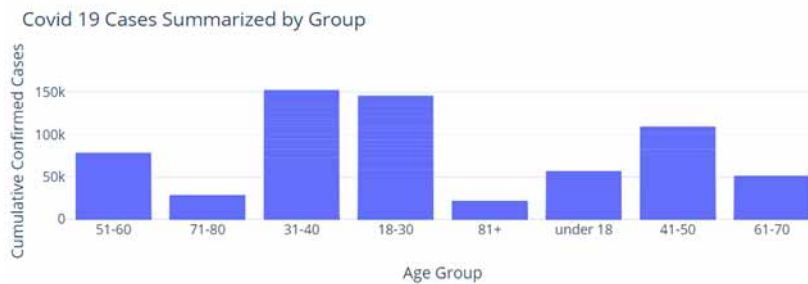


Figure 8. Impact of Covid19 on different age group as per timeliness (Source: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8UTBVA>)



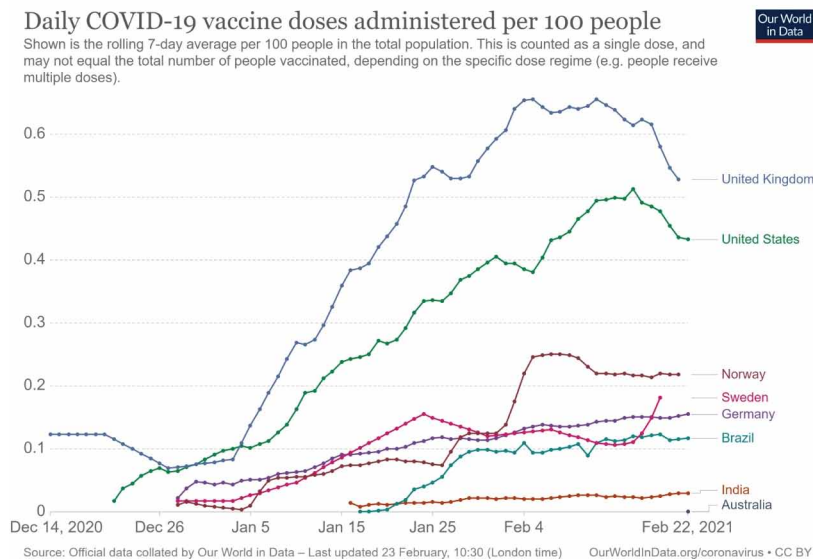
whole-body” side effects such as fatigue, headache and shivers after their first dose, compared with one fifth (19 percent) of patients who had not had COVID [46].

## CONCLUSION

In this paper, IoT based health care system is presented to alleviate the problems by chronic diseases.

The presented ontology-based healthcare architecture will be benefitted older people as they are facing challenges of loneliness, depression, physical disability, and weak immunity. They can improve their quality of life and monitor regularly their health beyond the doctor’s office. Healthcare providers

Figure 9. Specific vaccination dose regime [January-February 20,2021]



will get real-time data of a patient's health status and advise them accordingly and send notification and alerts. To fetch smart data in less time from web, ontology-based healthcare systems are most preferred. In the future, this work will be enhanced by incorporating an IoT based care plan features for monitoring a person's health conditions as well as medications too. Elderly people will continuously receive information from healthcare providers for overcoming their physical and hypertension issues through yoga, exercises, and meditation guidelines. Excessive intake of medicines may generate issues of multi-organ failure in old age people, so guidelines to physical exercise will be a source of motivation for them. This system will also keep track of emergency contacts and information serious patients on a timely basis. In the case of inconsequence, information of patients would be updated by health care providers to nearby healthcare so that treatment can be done without delay. This concept will be helpful for elderly patients who are living alone but will receive appropriate care in an emergency situation.

## REFERENCES

W3. (n.d.). *Semantic Web*. <https://www.w3.org/standards/semanticweb/>

Abinayaa, B., & Raja, A. (2016). Smart portable monitoring device for asthma patients. *Middle East Journal of Scientific Research*, 24(S1), 136–142.

Al-Khafajiy, M., Baker, T., Chalmers, C., Asim, M., Kolivand, H., Fahim, M., & Waraich, A. (2019). Remote health monitoring of elderly through wearable sensors. *Multimedia Tools and Applications*, 78(17), 24681–24706. doi:10.1007/s11042-018-7134-7

Alamri, A. (2018). Ontology middleware for integration of IoT healthcare information systems in EHR systems. *Computers*, 7(4), 51. doi:10.3390/computers7040051

Basenaru, L., Dobre, C., Ciobanu, R. I., & Balog, A. (2019, November). Patient Profile Using Ontologies in an Older Adults Monitoring IoT-Based Platform. In 2019 E-Health and Bioengineering Conference (EHB) (pp. 1-4). IEEE. doi:10.1109/EHB47216.2019.8970027

Basanta, H., Huang, Y. P., & Lee, T. T. (2016, April). Intuitive IoT-based H2U healthcare system for elderly people. In 2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC) (pp. 1-6). IEEE. doi:10.1109/ICNSC.2016.7479018

Bhagwat, N., Viviano, J. D., Voineskos, A. N., & Chakravarty, M. M. (2018). Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Computational Biology*, 14(9), e1006376. doi:10.1371/journal.pcbi.1006376 PMID:30216352

Bioportal. (n.d.). *COVID-19*. <https://bioportal.bioontology.org/ontologies/COVID-19>

COVID. (n.d.). <https://covid.joinzoe.com/post/vaccine-after-effects-more-common-in-those-who-already-had-covid>

Cruz, I. F., & Xiao, H. (2005). The role of ontologies in data integration. *Engineering Intelligent Systems for Electrical Engineering and Communications*, 13(4), 245.

Darwish, A., & Hassanien, A. E. (2011). Wearable and implantable wireless sensor network solutions for healthcare monitoring. *Sensors (Basel)*, 11(6), 5561–5595. doi:10.3390/s110605561 PMID:22163914

Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., & Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of COVID-19. *PLoS One*, 15(6), e0235187. doi:10.1371/journal.pone.0235187 PMID:32589673

Hassen, H. B., Dghais, W., & Hamdi, B. (2019). An E-health system for monitoring elderly health based on Internet of Things and Fog computing. *Health Information Science and Systems*, 7(1), 24. doi:10.1007/s13755-019-0087-z PMID:31695910

Hossain, M. S., & Muhammad, G. (2016). Cloud-assisted industrial internet of things (iiot)-enabled framework for health monitoring. *Computer Networks*, 101, 192–20. doi:10.1016/j.comnet.2016.01.009

Hosseinzadeh, M., Koohpayehzadeh, J., Bali, A. O., Asghari, P., Souri, A., Mazaherinezhad, A., & Rawassizadeh, R. et al. (2020). A diagnostic prediction model for chronic kidney disease in internet of things platform. *Multimedia Tools and Applications*, 1–18. doi:10.1007/s11042-020-09049-4

Islam, S. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K. S. (2015). The internet of things for health care: A comprehensive survey. *IEEE Access: Practical Innovations, Open Solutions*, 3, 678–708. doi:10.1109/ACCESS.2015.2437951

Jangra, P., & Gupta, M. (2018, August). A design of real-time multilayered smart healthcare monitoring framework using IoT. In 2018 International Conference on Intelligent and Advanced System (ICIAS) (pp. 1-5). IEEE. doi:10.1109/ICIAS.2018.8540606

Jara, A. J., Zamora, M. A., & Skarmeta, A. F. (2011). An internet of things-based personal device for diabetes therapy management in ambient assisted living (AAL). *Personal and Ubiquitous Computing*, 15(4), 431–440. doi:10.1007/s00779-010-0353-1

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2

Khalil, T., Khalid, S., & Syed, A. M. (2014, August). Review of Machine Learning techniques for glaucoma detection and prediction. In *2014 Science and Information Conference* (pp. 438-442). IEEE. doi:10.1109/SAI.2014.6918224

Kouris, I., & Koutsouris, D. (2014). Identifying risky environments for COPD patients using smartphones and internet of things objects. *International Journal of Computational Intelligence Studies*, 3(1), 1–17. doi:10.1504/IJCISTUDIES.2014.058642

Kumar, V. (2015). Ontology based public healthcare system in Internet of Things (IoT). *Procedia Computer Science*, 50, 99–102. doi:10.1016/j.procs.2015.04.067

Lal, P., Kumar, A., Kumar, S., Kumari, S., Saikia, P., Dayanandan, A., Adhikari, D., & Khan, M. L. (2020). The dark cloud with a silver lining: Assessing the impact of the SARS COVID-19 pandemic on the global environment. *The Science of the Total Environment*, 732, 139297. doi:10.1016/j.scitotenv.2020.139297 PMID:32408041

Lee, H., Park, Y. R., Kim, H. R., Kang, N. Y., Oh, G., Jang, I. Y., & Lee, E. (2020). Discrepancies in Demand of Internet of Things Services Among Older People and People With Disabilities, Their Caregivers, and Health Care Providers: Face-to-Face Survey Study. *Journal of Medical Internet Research*, 22(4), e16614. doi:10.2196/16614 PMID:32293575

Li, C., Hu, X., & Zhang, L. (2017). The IoT-based heart disease monitoring system for pervasive healthcare service. *Procedia Computer Science*, 112, 2328–2334. doi:10.1016/j.procs.2017.08.265

Ma, X., Wang, Z., Zhou, S., Wen, H., & Zhang, Y. (2018, June). Intelligent healthcare systems assisted by data analytics and mobile computing. In *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)* (pp. 1317-1322). IEEE. doi:10.1109/IWCMC.2018.8450377

Mae, J., Oey, E., & Kristiady, F. S. (2020). IoT based body weight tracking system for obese adults in Indonesia using realtime database. *E&ES*, 426(1), 012143. doi:10.1088/1755-1315/426/1/012143

Maglogiannis, I. (2009). Introducing intelligence in electronic healthcare systems: state of the art and future trends. In *Artificial Intelligence An International Perspective* (pp. 71–90). Springer. doi:10.1007/978-3-642-03226-4\_5

Mezghani, E., Exposito, E., Drira, K., Da Silveira, M., & Pruski, C. (2015). A semantic big data platform for integrating heterogeneous wearable data in healthcare. *Journal of Medical Systems*, 39(12), 185. doi:10.1007/s10916-015-0344-x PMID:26490143

Mora, H., Gil, D., Terol, R. M., Azorín, J., & Szymanski, J. (2017). An IoT-based computational framework for healthcare monitoring in mobile environments. *Sensors (Basel)*, 17(10), 2302. doi:10.3390/s17102302 PMID:28994743

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN Computer Science*, 2(1), 1-13.

Narin, A., Kaya, C., & Pamuk, Z. (2020). *Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks*. arXiv preprint arXiv:2003.10849

Onasanya, A., & Elshakankiri, M. (2019). Smart integrated IoT healthcare system for cancer care. *Wireless Networks*, 1–16. doi:10.1007/s11276-018-01932-1

Park, S. J., Subramaniam, M., Kim, S. E., Hong, S., Lee, J. H., Jo, C. M., & Seo, Y. (2017). Development of the elderly healthcare monitoring system with IoT. In *Advances in Human Factors and Ergonomics in Healthcare* (pp. 309–315). Springer. doi:10.1007/978-3-319-41652-6\_29

Pillai, S., Siddika, N., Apu, E. H., & Kabir, R. (2020). COVID-19: Situation of European Countries so Far. *Archives of Medical Research*.

Pinto, S., Cabral, J., & Gomes, T. (2017, March). We-care: An IoT-based health care system for elderly people. In *2017 IEEE International Conference on Industrial Technology (ICIT)* (pp. 1378-1383). IEEE. doi:10.1109/ICIT.2017.7915565

Pramanik, P. K. D., Pal, S., & Mukhopadhyay, M. (2019). Healthcare big data: A comprehensive overview. In *Intelligent Systems for Healthcare Management and Delivery* (pp. 72–100). IGI Global.



Ray, P. P. (2014, November). Home Health Hub Internet of Things (H 3 IoT): An architectural framework for monitoring health of elderly people. In *2014 International Conference on Science Engineering and Management Research (ICSEMR)* (pp. 1-3). IEEE. doi:10.1109/ICSEMR.2014.7043542

Recalcati, S. (2020). Cutaneous manifestations in COVID-19: a first perspective. *Journal of the European Academy of Dermatology and Venereology*.

Rodriguez-Morales, A. J., Bonilla-Aldana, D. K., Tiwari, R., Sah, R., Rabaan, A. A., & Dhama, K. (2020). COVID-19, an emerging coronavirus infection: Current scenario and recent developments-an overview. *Journal of Pure & Applied Microbiology*, 14(1), 6150. doi:10.22207/JPAM.14.1.02

Shah, S. T. U., Badshah, F., Dad, F., Amin, N., & Jan, M. A. (2019). Cloud-assisted IoT-based smart respiratory monitoring system for asthma patients. In *Applications of Intelligent Technologies in Healthcare* (pp. 77–86). Springer. doi:10.1007/978-3-319-96139-2\_8

Szilagyi, I., & Wira, P. (2016, October). Ontologies and Semantic Web for the Internet of Things-a survey. In *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society* (pp. 6949-6954). IEEE. doi:10.1109/IECON.2016.7793744

Tun, S. Y. Y., Madanian, S., & Mirza, F. (2020). Internet of things (IoT) applications for elderly care: A reflective review. *Aging Clinical and Experimental Research*, 1–13. doi:10.1007/s40520-020-01545-9 PMID:32277435

Tyagi, S., Agarwal, A., & Maheshwari, P. (2016, January). A conceptual framework for IoT-based healthcare system using cloud computing. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)* (pp. 503-507). IEEE. doi:10.1109/CONFLUENCE.2016.7508172

Wang, X., Wang, Z., Weng, J., Wen, C., Chen, H., & Wang, X. (2018). A new effective machine learning framework for sepsis diagnosis. *IEEE Access: Practical Innovations, Open Solutions*, 6, 48300–48310. doi:10.1109/ACCESS.2018.2867728

WHO. (2021). Status. [https://extranet.who.int/pqweb/sites/default/files/documents/Status\\_COVID\\_VAX\\_16Feb2021.pdf](https://extranet.who.int/pqweb/sites/default/files/documents/Status_COVID_VAX_16Feb2021.pdf)

Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining Electronic Health Records (EHRs) A Survey. *ACM Computing Surveys*, 50(6), 1–40. doi:10.1145/3127881

*Sakshi Gupta has received her Bachelor of Computer Applications and Master of Computer Applications from Uttar Pradesh Technical University (Now Dr APJ AKTU), Lucknow, Uttar Pradesh. She was amongst TOP2 Meritorious student of MCA and had won Scholarship. Currently, She is doing Ph.D. from Birla Institute of technology Mesra, Ranchi. She has experience of more than 3 years in Academics. She has keen interest in the area of Internet of Things, Mobile Ad hoc network, and Wireless Sensor networks. She has published about 5 research Papers in reputed Conferences at national/international level indexed in including ESCI, Scopus, Google Scholar. She has also attended various technical Conferences/Workshops/Faculty Development Programmes/Orientation Programmes from organization repute.*

*Umang Singh (PhD) is presently working as Associate Professor in Institute of Technology & Science, Ghaziabad. She is renowned for her keen interest in the area of Mobile Networks, IoT, Data Analytics and Machine learning. Dr. Umang has experience of more than 17 Years in Academics. She is involved in active research and has been guiding M.Tech, Ph.D. students. She has published about 80+ research Papers in reputed Journals & Conferences indexed in including SCI, ESCI, SCIE, Scopus, Google Scholar, Thomas Reuters, DBLP credited to her name. She served as Guest Editor for special issues of journals which include "International Journal of e-Collaborations"(IGI Global, USA, 2020), American Journal of Artificial Intelligence(SciencePG, NY, USA, 2020) and International Journal of Information Technology (BJIT 2010) Special Issue and edited 6 Proceedings, 3 Souvenir and 4 Books. Dr Umang is Board of Referees Springer IJIT(Scopus Indexed), Editorial Board Member Inderscience IJFSE(Switzerland) and Technical Programme Committee of national/international journals/Conferences/Seminars repute. Dr. Umang has also delivered session as Keynote Speaker/chaired/conducted a Technical Sessions in Conferences/Workshops/Faculty Development Programmes/Orientation Programmes repute. Dr Umang is Senior Member of IEEE, Member of Computer Society of India, IAENG, MIR Labs etc. Dr. Umang has been delivering invited talks, guest talks at prominent places and organizations including Indian Air Force. Dr. Umang has received Certificate of Appreciation from VSM, Airforce Station Hindan, Ghaziabad (UP). Dr. Umang has received "Young Active Member award" for year 2007-2008 from Computer Society of India. She has also received "Young Faculty in Science" in Year 2017 from VIFA, Chennai.*

# Improvement in Task Scheduling Capabilities for SaaS Cloud Deployments Using Intelligent Schedulers

Supriya Sawwashere, Kalinga University, India

## ABSTRACT

Task scheduling on the cloud involves processing a large set of variables from both the task side and the scheduling machine side. This processing often results in a computational model that produces efficient task-to-machine maps. The efficiency of such models is decided based on various parameters like computational complexity, mean waiting time for the task, effectiveness to utilize the machines, etc. In this paper, a novel Q-Dynamic and Integrated Resource Scheduling (DAIRS-Q) algorithm is proposed which combines the effectiveness of DAIRS with Q-Learning in order to reduce the task waiting time and improve the machine utilization efficiency. The DAIRS algorithm produces an initial task-to-machine mapping, which is optimized with the help of a reward and penalty model using Q-Learning, and a final task-machine map is obtained. The performance of the proposed algorithm showcases a 15% reduction in task waiting time and a 20% improvement in machine utilization when compared to DAIRS and other standard task-scheduling algorithms.

## KEYWORDS

Cloud Deployments, Cloud-Based IoT, DAIRS, Intelligent Schedulers, Machine Mapping, Machine Utilization, Q-Learning, Task Scheduling, Task Waiting Time

## 1. INTRODUCTION

Scheduling tasks over the cloud is a multi-domain problem, which includes pattern analysis, filtering, classification, clustering and prediction. Usually the following processes are followed in order to schedule cloud tasks (Asghari et al., 2020),

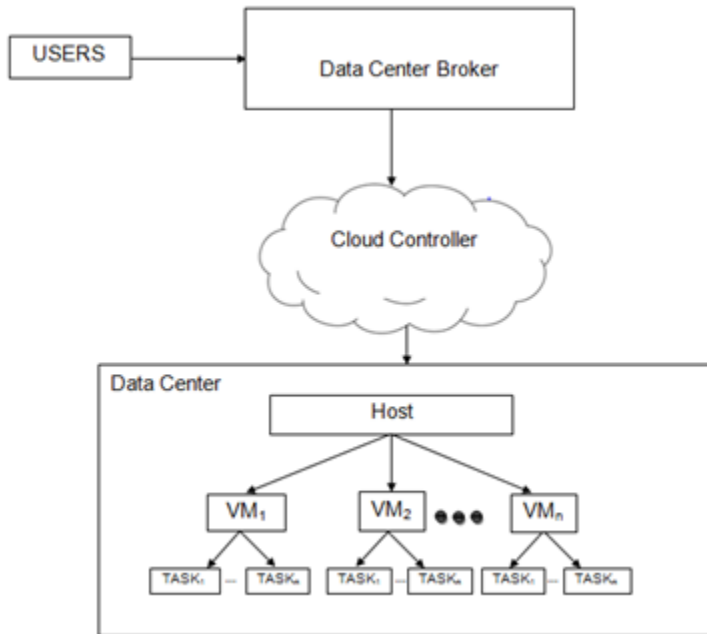
- Identification of undertaking boundaries from the task dataset
- Identification of machine parameters from the asset pool
- Strategizing rules and thresholds for machine and task scheduling
- Task execution on the given machine
- Evaluation of irregularities in execution, and changing methodology based to output parameters
- Post processing of tasks and machines if needed

In view of these steps, the researchers can see that at first the task parameters must be investigated. These parameters must incorporate essential assignment measurements like the undertaking execution delay, the task cutoff time, the task holding up time, while they can likewise incorporate optional

DOI: 10.4018/IJBDAH.287104

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Figure 1. A typical task scheduler



parameters like undertaking mutual exclusiveness, shared reliance, and others (Nawrocki & Sniezynski, 2020). Typically, the all-out assignment execution prerequisites are administered by condition 1,

$$T_{TE} = F(T_{ed}, T_{dl}, T_{wt}, [T_{sec}]) \dots (1)$$

Where,  $T_{TE}$  is the total task execution requirement,  $T_{ed}$  is the total task execution delay,  $T_{dl}$  is the task deadline,  $T_{wt}$  is the task waiting time, while,  $T_{sec}$  are secondary application specific parameters needed to execute the task.

Once the task parameters are identified, then the resource parameters are observed, and evaluated. These parameters are again divided into primary and secondary parameters. Primary parameters include but are not limited to number of execution units available, capacity of each unit to execute the task, execution requirements for the resources, and others (Sui et al., 2019). A typical task scheduler can be observed from figure 1, wherein the tasks coming from users are given to the data center broker, the broker sends these tasks to the cloud controller for processing. The controller finally gives it to the host for further processing and scheduling on different machines.

The next section describes about such task scheduling systems in brief, and is followed by the proposed DAIRS-Q algorithms. This text further evaluates the said algorithm on different application specific datasets, and compares its efficiency with some state-of-the-art methods. Finally, the concludes with some interesting observations about the proposed protocol, and recommendations on how to further explore the field of work.

## 2. LITERATURE REVIEW

AI has become a true norm for task scheduling research. The work (Ge & Liu, 2020) uses Q-Learning for scheduling undertakings on a cloud-based Internet of Things (IoT) climate. The principle bit of leeway of Q-Learning is that, it gives a motivating force (reward) and punishment procedure while

finding out about the basic issue. For example, in Ge & Liu, (2020), analysts have utilized a worldwide view strategy for task scheduling, wherein the calculation's choice is compensated if the general cloud execution improved by it, or it is given a punishment if the cloud execution debases. The prize and punishment is as a mathematical worth, which is augmented by the calculation. They have isolated the errand nodes into energy unwinding and energy tense nodes. The energy tense nodes are answerable for imparting back and forth between the cloud and the clients. While the energy unwinding node assumes control over a portion of the heap structure the energy tense nodes. The principle point of the calculation is to improve the network lifetime, by enhancing the heap on every one of the node. A similar calculation can be applied to any sort of assignment scheduling applications, and it is prescribed that perusers apply it to assess its exhibition on various applications. The calculation can be stretched out by supplanting Q-Learning with fortification realizing, which is an unrivaled method for quicker union and better learning results. This can be seen from the work in Melnik & Nasonov, (2019), wherein fortification learning is joined with neural networks for scheduling work processes. The engineering takes in the assignments which need scheduling, and offers them to a processing climate. The figuring climate assesses a prize capacity dependent on the scheduling done by the neural network. Normally the neural network plans errands arbitrarily, and attempts to limit the blunder in scheduling with the assistance of learning models like Levenberg–Marquardt, inclination plummet, and so forth Support learning can be broadened utilizing a more intricate preparing condition set, which can think about both essential and auxiliary boundaries for undertakings and execution units the same. For example, the work in Waschneck et al., (2018), broadens support learning, and assesses a profound fortification learning component dependent on Deep Mind which is Google's Deep Q-Learning Network (DQN). It utilizes a mix of Markov learning measure with directed learning (so the calculation find out about the execution climate), for better undertaking scheduling effectiveness. The model can improve the asset use, and decrease the undertaking holding up time, yet has a high displaying multifaceted nature. Since each time something changes in the processing plant climate, a comparative change must be made in the computerized twin climate. Generally, ongoing undertaking execution frameworks have an enormous number of changes happening occasionally, subsequently this system isn't appropriate for such exceptionally powerful conditions. However, the exploration accomplished for planning a 2-level learning calculation is excellent, and hence is utilized as the reason for this fundamental examination.

Another Q-Learning based calculation is referenced in Kim et al., (2019), wherein an IoT network comprising of temperature, mugginess and weight sensors is thought of. Here, a MDP or Markov choice cycle is applied, which totals the prizes given to every arrangement set to locate the last prize worth. The work in Waschneck et al., (2018) and Kim et al., (2019) is stretched out in Zhang et al., (2019), wherein a profoundly dispersed climate is considered for task scheduling in programming as a help (SaaS) cloud. They have utilized the idea of versatile unique programming (ADP) to plan undertakings dependent on both assignment and execution unit boundaries. It tends to be seen that the proposed component performs in a way that is better than a portion of the cutting-edge calculations, and can be utilized as a decent sending methodology when scheduling assignments in a conveyed climate.

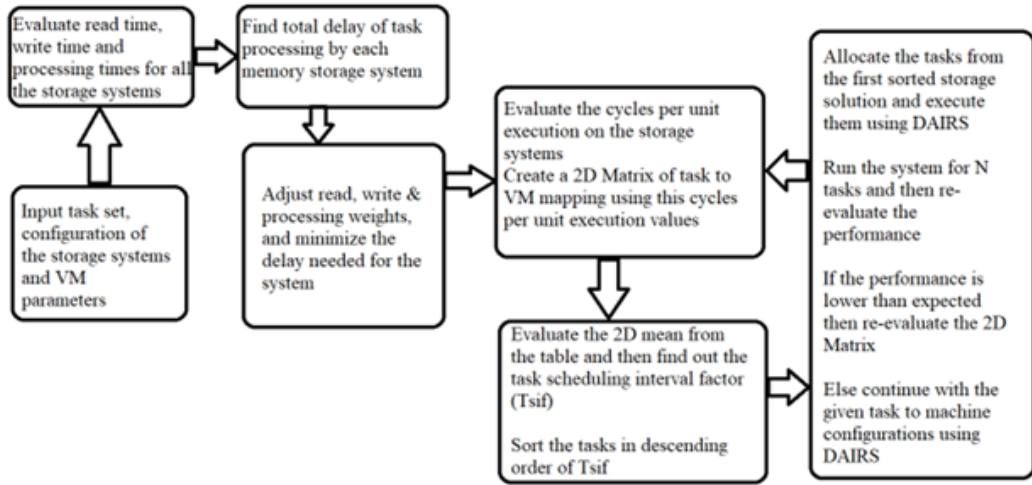
A use of this calculation can be seen from Hu et al., (2019), wherein the errand scheduling issue is first changed over into min-max number direct programming (MMILP), and afterward it is changed over to a distinguishable curved target work with a unimodular imperative grid for adding non-linearity. The capacity and the framework information gain from the conveyed disseminated network, and improve the undertaking scheduling productivity by diminishing the computational deferral, and lessening the quantity of bytes moved during scheduling. The work in Nawrocki et al., (2020) presents an energy viable arrangement, uses energy relevant boundaries like CPU use, network association, defer required for calculations, and so on to assess the best scheduling procedure. A similar setting mindfulness can be utilized to upgrade some other boundary, similar to defer required for scheduling, mean holding up time, cutoff time hit proportion, and so forth Because of which this work has been chosen to advance certain boundaries in the fundamental exploration. The

energy utilization is diminished with the assistance of neighborhood learning, and the conveyed task scheduler can plan the assignments on the nodes that are the most energy proficient. This decreases the framework's productivity to enhance other undertaking and node boundaries, yet improves the general framework lifetime. Such a framework can be utilized where the undertaking to node scheduling needs to streamline just a single boundary. In any case, as number of boundaries increment, the setting of the issue will in general change, and complex calculations like Stavrinides & Karatza, (2017), that utilizations rough calculations for task-scheduling can be used. Because of this, the settle on range and administration level understanding infringement proportion lessens, accordingly improving the errand scheduling proficiency.

Another CNN roused cross breed profound neural network scheduler is talked about in Zang et al., (2019), which uses convolution two-dimensional change technique to perform task scheduling. Another 2-stage task scheduler is portrayed in Zhang & Zhou, (2017), which uses both current and recorded information for scheduling errands. Because of the utilization of both authentic and current errand and node information, the calculation can foresee task examples and VM utility examples. A utilization of this framework can be seen in Zheng et al., (2020), wherein task scheduling is applied to savvy city situations. They presume that Extended Hungarian calculation with round support line (EHGC) is a superior calculation when contrasted and FIFO and other shortsighted frameworks. The use of profound learning and fortification learning can be tried for savvy urban communities. Scheduling can likewise be conveyed as a cloud administration, the work in Moorthy & Pabitha, (2019) features a utilization of sending scheduling calculations on the cloud, and giving it as assistance. It tends to be utilized as a strong use-case for planning scheduling calculations.

An epic particle swarm optimization (PSO) calculation for task scheduling is depicted in Ebadifard & Babamir, (2017), wherein the wellness work is changed to be a mix of make-range and the asset use. A Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) calculation is applied in Khorsand & Ramezanpour, (2020), which uses the best-most exceedingly terrible strategy for task scheduling. It additionally utilizes the oversimplified approach as utilized in Dong et al., (2019), and can accomplish comparative execution like Dong et al., (2019), but on the other hand can decrease the energy utilization because of the utilization of energy as a wellness boundary. However, both Dong et al., (2019) and Prasanna Kumar & Kousalya, (2019) have restricted use cases because of the restricted capacity of the calculation demonstrated, which can be improved by the work in Shobha Rani & Pounambal, (2019). Here, another profound fortification learning calculation is depicted that uses numerous errand and node boundaries for better optimization. Like PSO (Dong et al., 2019), a crow inquiry optimization (CSO) is portrayed in Zhou et al., (2018). It utilizes comparable exploration boundaries like Dong et al., (2019), and accomplishes a comparative execution on shortsighted datasets. The work can be streamlined utilizing a profound learning network as depicted in Peng et al., (2019), wherein task offloading is improved preparing. It utilizes profound learning CNN for task scheduling, and gives better make range and better computational usage when contrasted and Zhou et al., (2018). A comparative way to deal with Zhou et al., (2018) is referenced in Huang et al., (2019), where an adjusted hereditary calculation (GA) joined with voracious methodology (MGGS) is applied. The GA utilizes a voracious methodology for wellness assessment and traverse activities, which makes it viable for scheduling undertakings. However, like Dong et al., (2019) and Zhou et al., (2018), the methodology has characteristic downsides because of set number of boundary use for wellness assessment. Accordingly, the work in Mostafavi & Hakami, (2020), which utilizes a profound Q-Learning network can be utilized for execution upgrade. Comparative profound learning techniques are referenced in Huang et al., (2019), Mostafavi & Hakami, (2020) and Tong et al., (2019), where Q-Learning and fortification learning is either utilized independently or joined with neural networks to accomplish unrivaled execution. Consequently, the underlying research likewise utilizes Q-Learning in mix with DAIRS to improve the framework execution for load adjusting over the cloud.

Figure 2. Proposed DAIRS-Q method for real time clouds



### 3. PROPOSED DAIRS-Q ALGORITHM

The original DAIRS algorithm (Anjum et al., 2020) does not perform well under hybrid storage conditions, and nowadays all cloud systems are based on hybrid storage (which combine SSDs, HDDs, etc.). Therefore, the proposed Q-learning based DAIRS is proposed that uses dummy reads and writes in order to evaluate the performance of the original algorithm, and get the final optimized scheduling results. The overall architecture of the proposed DAIRS can be observed from figure 2. The proposed hybrid DAIRS (DAIRS-Q) system, works in the following steps,

1. For a cloud with ‘N’ types of storage systems, arrange each of these systems in descending order of processing. Due to this the highest performance system is placed at the top, while the lowest performing system is placed at the bottom.
2. Perform ‘N’ dummy reads and ‘N’ dummy writes on each of the cloud systems, and observe the following parameters,
  - a. The reading delay for all systems  $Tr_1, Tr_2, \dots, Tr_N$
  - b. The writing delay for all systems  $Tw_1, Tw_2, \dots, Tw_N$
  - c. The processing delay for all systems  $Tp_1, Tp_2, \dots, Tp_N$
3. Evaluate the total time needed for processing one task using the following equation 1,

$$Td_i = w_r * Tr_i + w_w * Tw_i + w_p * Tp_i \quad (1)$$

where,  $w_r$ ,  $w_w$  and  $w_p$  are the weights for reading, writing and processing. Higher values of weights indicate that the application needs that particular element to be enhanced. For instance, an application that requires faster processing will have higher value of  $w_p$ , while the values of  $w_r$  and  $w_w$  would be lower than that of  $w_p$ .

**Table 1. Relationship between system memory type and input task**

$Tn1$ $Tl$	$Tn2$ $Tl$	$Tn3$ $Tl$	....	$TnN$ $Tl$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$Tn1$ $Tk$	$Tn2$ $Tk$	$Tn3$ $Tk$	...	$TnK$ $Tk$

- Evaluate the lowest total completion time for each of these machines and store them into an array. Let the elements in this array be named as D1, D2, D3, ... DN
- For each storage system, the time needed to process the task on the system will be different. Let Tni represent the time needed for a task to be executed on that system, this can be evaluated with the help of equation 2 as follows,

$$Tni = \frac{Di * Ni}{Tdi} \quad (2)$$

where Tdi is the sum of processing time for the cloud, Ni is the total number of task units to be executed on the cloud, Di is the per unit task delay for the system.

- Now evaluate the following dependency table, which indicates the execution time for a given task on a given set of machines,
- This table consists of all 'k' tasks on the rows, and all the 'N' machine configurations on the column side
- Find the double average value of the task execution threshold using the following equation 3,

$$Sth = \frac{\sum_{i=1}^N \sum_{j=1}^k Tni Tj}{N * k} \quad (3)$$

- Now find the interval factor for the task using the following formula,

$$IF_t = \frac{Sth}{N} \quad (4)$$

Here, 'N' is used as the number of memory configurations are 'N' in the system.

10. Find the column-wise average value from the dependency table, and let the values of these averages be  $AT_1, AT_2, \dots, AT_k$
11. Arrange these tasks in descending order of the AT values, and let the new order for these tasks be  $Ts_1, Ts_2, \dots, Tsk$
12. For allocating a task 'i' to any machine, it must fulfill the given condition,

$$(i-1)*IF_i \leq AT_d \leq i*IF_i \quad (5)$$

Where, 'd' is the ID of the task

Based on the equation 5, the task is given a penalty value if it doesn't follow equation 5, while it is given a reward value if the task follows it. All tasks are allocated to a machine only if they are rewarded, else the task is left unallocated. The process is repeated unless all tasks are successfully allocated by the system. Based on this allocation, the task with higher requirement is given to a memory system with better performance, while a task with lower requirement is given to a memory system with lower configuration/performance. This enables the task scheduler to schedule tasks with higher efficiency, and get better quality of experience performance for the end user. This performance can be observed from the next section, where a statistical comparison is made between the existing DAIRS and IDE algorithms with the proposed DAIRS-Q algorithm.

#### 4. RESULT EVALUATION AND COMPARISON

To compare the performance of the proposed DAIRS-Q algorithm with the existing IDE (Wu et al., 2018) and DAIRS algorithms, the NASA load balancing dataset from the following website is used, [https://www.cs.huji.ac.il/labs/parallel/workload/l\\_nasa\\_ipsc/index.html](https://www.cs.huji.ac.il/labs/parallel/workload/l_nasa_ipsc/index.html)

This dataset consists of more than 100k records with an average makespan of 600 ms for each task. Makespan is the average delay needed for execution of the task. Based on this dataset, the average cloud utilization ratio (CUR), the delay needed for task scheduling (Da) and the mean task waiting time (Tmwt) are evaluated. The following tables indicate the comparison of these values for different algorithms.

Similarly, the other parameters can be observed from table 2 and 3 as follows:

It can be observed that the proposed DAIRS-Q reduces the delay of task scheduling by 15% when compared with the existing DAIRS and IDE algorithms, while the cloud utilization ratio is improved by more than 10%. This happens due to the dummy reads and writes architecture which is proposed, that allows the system to evaluate the performance of the system before deployment, and therefore assigns the best task scheduling plan in place for the given set of tasks and virtual machine combinations.



Table 2. Comparison of cloud utilization ratio for different algorithms

VM-s	Tasks	CUR (DAIRS)	CUR (IDE)	CUR (DAIRS-Q)
10	1000	74.50	70.95	80.81
10	2000	75.60	72.00	82.00
10	5000	75.90	72.29	82.33
10	10000	76.30	72.67	82.76
10	20000	77.50	73.81	84.06
10	50000	79.10	75.33	85.80
10	100000	79.90	76.10	86.66
20	1000	80.51	76.68	87.33
20	2000	81.40	77.52	88.29
20	5000	82.29	78.37	89.25
20	10000	83.17	79.21	90.21
20	20000	84.06	80.05	91.17
20	50000	84.94	80.90	92.13
20	100000	85.83	81.74	93.09
50	1000	86.71	82.59	94.06
50	2000	87.60	83.43	95.02
50	5000	88.49	84.27	95.98
50	10000	89.37	85.12	96.94
50	20000	90.26	85.96	97.90
50	50000	91.14	86.80	98.86
50	100000	92.03	87.65	99.82

**Table 3. Comparison of task execution delay**

VM-S	Tasks	Da (s) (DAIRS)	Da (s) (IDE)	Da (s) (DAIRS-Q)
10	1000	5.20	4.95	4.13
10	2000	5.50	5.24	4.37
10	5000	5.70	5.43	4.53
10	10000	5.90	5.62	4.68
10	20000	6.23	5.93	4.94
10	50000	7.10	6.76	5.63
10	100000	7.50	7.14	5.95
20	1000	2.60	2.48	2.07
20	2000	2.75	2.62	2.18
20	5000	2.85	2.71	2.26
20	10000	2.95	2.81	2.34
20	20000	3.12	2.97	2.48
20	50000	3.55	3.38	2.82
20	100000	3.75	3.57	2.98
50	1000	1.30	1.24	1.03
50	2000	1.38	1.31	1.09
50	5000	1.43	1.36	1.13
50	10000	1.48	1.40	1.17
50	20000	1.56	1.48	1.23
50	50000	1.78	1.69	1.41
50	100000	1.88	1.79	1.49

**Table 4. Comparison of mean waiting time for tasks**

VM-S	Tasks	Tmwt (s) DAIRS)	Tmwt (s) (IDE)	Tmwt (s) (DAIRS-Q)
10	1000	6.90	6.57	6.12
10	2000	7.50	7.14	6.66
10	5000	8.90	8.48	7.90
10	10000	13.50	12.86	11.98
10	20000	14.90	14.19	13.22
10	50000	15.60	14.86	13.84
10	100000	16.20	15.43	14.38
20	1000	3.45	3.29	3.06
20	2000	3.75	3.57	3.33
20	5000	4.45	4.24	3.95
20	10000	6.75	6.43	5.99
20	20000	7.45	7.10	6.61
20	50000	7.80	7.43	6.92
20	100000	8.10	7.71	7.19
50	1000	1.73	1.64	1.53
50	2000	1.88	1.79	1.66
50	5000	2.23	2.12	1.97
50	10000	3.38	3.21	3.00
50	20000	3.73	3.55	3.31
50	50000	3.90	3.71	3.46
50	100000	4.05	3.86	3.59

## 5. CONCLUSION AND FUTURE SCOPE

Based on the result evaluation it can be observed that the proposed algorithm outperforms both DAIRS and IDE algorithms in terms of cloud utilization ratio, delay of task execution and mean task waiting time. The cloud utilization is improved by almost 10%, while the delay for execution is reduced by 15, and the mean waiting delay is reduced by 10% as well. This indicates that the proposed algorithm can be applied to any real-time cloud deployments, thereby improving the overall applicability of the proposed DAIRS-Q system. The proposed algorithm can be further improved with the addition of more sophisticated scheduling algorithms that do not require dummy reads and writes, because using these dummy requests increases the computational overheads on the system. In order to reduce them, machine learning techniques like deep-reinforcement learning can be used, along with better mapping algorithms for scheduling tasks on real-time clouds.

## REFERENCES

- Anjum, Chaudhary, & Karanjekar. (2020). Dynamic Load Balancing Scheduling Algorithm for Cloud Data Centers. *International Research Journal of Modernization in Engineering Technology and Science*, 1252-1256.
- Asghari, A., Sohrabi, M. K., & Yaghmaee, F. (2020). Task scheduling, resource provisioning, and load balancing on scientific workflows using parallel SARSA reinforcement learning agents and genetic algorithm. *The Journal of Supercomputing*, 1–29.
- Dong, T., Xue, F., Xiao, C., & Li, J. (2019). Task scheduling based on deep reinforcement learning in a cloud manufacturing environment. Wiley Online Library.
- Ebadifard, F., & Babamir, S. M. (2017). A PSO-based task scheduling algorithm improved using a load balancing technique for the cloud computing environment. Wiley Online Library.
- Ge, J., & Liu, B. (2020). Q-learning based flexible task scheduling in a global view for the Internet of Things. Wiley Online Library.
- Hu, Li, & Luo. (2019). Time- and Cost- Efficient Task Scheduling across Geo-Distributed Data Centers. *IEEE Transactions on Parallel and Distributed Systems*, 705-718.
- 2Huang, J., Li, S., & Chen, Y. (2019). Revenue-optimal task scheduling and resource management for IoT batch jobs in mobile edge computing. *Peer-to-Peer Networking and Applications*, 1–12.
- Khorsand, R., & Ramezanpour, M. (2020). An energy-efficient task-scheduling algorithm based on a multi-criteria decision-making method in cloud computing. Wiley Online Library.
- Kim, D., Lee, T., Kim, S., Lee, B., & Youn, H. Y. (2019). Adaptive packet scheduling in IoT environment based on Q-learning. *Journal of Ambient Intelligence and Humanized Computing*, 1–11. doi:10.1007/s12652-019-01351-w
- Melnik, M., & Nasonov, D. (2019). Workflow scheduling using Neural Networks and Reinforcement Learning. *8th International Young Scientist Conference on Computational Science*, 29-36. doi:10.1016/j.procs.2019.08.126
- Moorthy & Pabitha. (2019). Optimal provisioning and scheduling of analytics as a service in cloud computing. Wiley Online Library.
- Mostafavi, S., & Hakami, V. (2020). A Stochastic Approximation Approach for Foresighted Task Scheduling in Cloud Computing. *Wireless Personal Communications*, 114(1), 1–25. doi:10.1007/s11277-020-07398-9
- Nawrocki, P., & Sniezynski, B. (2020). Adaptive Context-Aware Energy Optimization for Services on Mobile Devices with Use of Machine Learning. *Wireless Personal Communications*, 1–29.
- Nawrocki, P., Sniezynski, B., Kolodziej, J., & Szykiewicz, P. (2020). Adaptive context-aware energy optimization for services on mobile devices with use of machine learning considering security aspects. *20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, 708-717. doi:10.1109/CCGrid49817.2020.00-19
- Peng, Lin, Cui, Li, & He. (2019). A multi-objective trade-off framework for cloud resource scheduling based on the Deep Q-network algorithm. *Cluster Computing*, 1-15.
- Prasanna Kumar, K. R., & Kousalya, K. (2019). Amelioration of task scheduling in cloud computing using crow search Algorithm. *Neural Computing & Applications*, 1–7.
- Shobha Rani & Pounambal. (2019). Deep learning based dynamic task offloading in mobile cloudlet Environments. *Evolutionary Intelligence*, 1-9.
- Stavriniades, G. L., & Karatza, H. D. (2017). Scheduling real-time bag-of-tasks applications with approximate computations in SaaS clouds. Wiley Online Library.
- Sui, X., Li Li, D. L., Wang, H., & Yang, H. (2019). Virtual machine scheduling strategy based on machine learning algorithms for load balancing. *EURASIP Journal on Wireless Communications and Networking*, 2019(1), 1–16. doi:10.1186/s13638-019-1454-9
- Tong, Deng, Hongjian, & Liu. (2019). QL-HEFT: a novel machine learning scheduling scheme base on cloud computing environment. *Neural Computing and Applications*, 1-18.

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2

Waschneck, B., Reichstaller, A., Belzner, L., Altenmuller, T., Bauernhansl, T., Knapp, A., & Kyek, A. (2018). Optimization of Global Productin Scheduling with Deep Reinforcement Learning. *51st CIRP Conference on Manufacturing Systems, ScienceDirect*, 1264-1269.

Wu, Liu, & Zhao. (2018). An improved differential evolution algorithm for solving a distributed assembly flexible job shop scheduling problem. *Memetic Computing*, 1-21.

Zang, Wang, Song, Lu, Li, Wang, & Zhao. (2019). Hybrid Deep Neural Network Scheduler for Job-Shop Problem Based on Convolution Two-Dimensional Transformation. *Computational Intelligence and Neuroscience*, 1-20.

Zhang, Ma, Xiao, Li, & Lin. (2019). *Two-level task scheduling with multi-objectives in geo-distributed and large-scale SaaS cloud*. Springer Nature.

Zhang, P. Y., & Zhou, M. C. (2017). Dynamic Cloud Task Scheduling Based on a Two-Stage Strategy. *IEEE Transactions on Automation Science and Engineering*, 1–12.

1Zheng, X., Li, M., & Guo, J. (2020). Task scheduling using edge computing system in smart city. Wiley Online Library.

Zhou, Li, Zhu, Xie, Abawajy, & Chowdhury. (2018). An improved genetic algorithm using greedy strategy toward task scheduling optimization in cloud environments. *Neural Computing and Applications*, 1-19.

*Supriya Sawwashere is pursuing Ph.D. in Computer Science Engineering from Kalinga University, Raipur. Currently she is working in J. D. College of Engineering & Technology, Nagpur. She has teaching experience of 13.5 years. She has completed his M.Tech. in Computer Science Engineering from G.H. Rasoni College of Engineering, Nagpur (Maharashtra), India. She has completed his B.E. in Information Technology. She has published 8 papers in international journals and 3 papers in international conferences. Her areas of research interest are network security, theory of computation, computer graphics, and IoT.*

# Table of Contents

## International Journal of Big Data and Analytics in Healthcare

Volume 6 • Issue 2 • July-December-2021 • ISSN: 2379-738X • eISSN: 2379-7371

### Open Access Article

- 1      **A Predictive Analytics Framework for Blood Donor Classification**  
Kavita Pabreja, Maharaja Surajmal Institute, GGSIP University, Delhi, India  
Akanksha Bhasin, GGSIP University, Delhi, India
- 15     **Different Approaches to Reducing Bias in Classification of Medical Data by Ensemble Learning Methods**  
Adem Doganer, Kahramanmaraş Sutcu Imam University, Kahramanmaraş, Turkey
- 31     **ICTs and Domestic Violence (DV): Exploring Intimate Partner Violence (IPV)**  
Bolanle A. Olaniran, Texas Tech University, USA
- 45     **Using Data Science Software to Address Health Disparities**  
Jose O. Huerta, University of North Texas, USA  
Gayle L. Prybutok, University of North Texas, USA  
Victor Prybutok, University of North Texas, USA
- 59     **Big Data Applications in Vaccinology**  
Joseph E. Kasten, Pennsylvania State University, York, USA


### COPYRIGHT

The **International Journal of Big Data and Analytics in Healthcare (IJBD AH)** (ISSN: 2379-738X; eISSN: 2379-7371), Copyright © 2021 IGI Global. From the journal's inception, January 1, 2016, to December 31, 2020, all rights, including translation into other languages is reserved by the publisher, unless otherwise stated in the article manuscript. As of January 1, 2021, this journal operates under the gold Open Access model, whereby all content published after this date is distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License (<http://creativecommons.org/licenses/by/4.0/>) where copyright for the work remains solely with the author(s) of the article manuscript. Product or company names used in this journal are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark. The views expressed in this journal are those of the authors but not necessarily of IGI Global.

The *International Journal of Big Data and Analytics in Healthcare* is indexed or listed in the following: ACM Digital Library; Cabell's Directories; Google Scholar

# A Predictive Analytics Framework for Blood Donor Classification

Kavita Pabreja, Maharaja Surajmal Institute, GGSIP University, Delhi, India

 <https://orcid.org/0000-0001-9856-0900>

Akanksha Bhasin, GGSIP University, Delhi, India

## ABSTRACT

India faces numerous challenges to meet ever-increasing demand of human blood so as to improve the health indicators across its rural and urban population. The gap between demand and supply can be fulfilled by increasing voluntary blood donations. Hence, it becomes important to understand the attitude of population towards blood donations. In this paper an effort has been made to identify features in order of their importance that affect the decision of a person to become a blood donor. This research uses extensive visualization techniques to get an insight into potential blood donor characteristics and then applies classification technique to classify youth of an Indian state university as donor or non-donor. The k-nearest neighbour classification algorithm discovers the relationship between attributes of blood donors and hence predicts the outcome. The important factors that dissuade potential donors from donating blood have been extracted that can be worked upon to meet the demand of blood to save human lives.

## KEYWORDS

Blood Donor, Classification, Data Visualization, K-Nearest Neighbour, Lazy Learner Algorithm, Logistic Regression, Machine Learning, Random Feature Elimination

## INTRODUCTION

Human blood is the precious constituent of life and there is no substitute for it. There has always been an acute shortage of human blood as far as a developing nation like India is concerned as stated by Verma et al. (2016). It is mentioned by Abolghasemi et al. (2009) that the rate of blood donations in developing countries is eighteen times lesser as compared to that of developed countries. Voluntary blood donations meet a significant portion of blood requirement in countries with higher income as explained by Nigatu and Demissie (2015). This non-remunerated donation has been considered as best and safest by Gharebghagian, 2005 and Rahman et al. (2011).

A report on National Estimation of Blood Requirement in India has mentioned that the country faces many challenges in maintaining a sufficient supply of blood and its products. With an ever increase in Indian population augmented by advancement in clinical medicine, the demand of blood far outweighs its supply. This is also emphasized by Agrawal et al. (2013) and Benedict et al. (2012)

DOI: 10.4018/IJBDAH.20210701.oa1

This article, published as an Open Access article on April 23, 2021 in the gold Open Access journal, International Journal of Big Data and Analytics in Healthcare (converted to gold Open Access January 1, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

in their research reports. According to World Health Organization (WHO), every country should be able to provide safe and adequate blood to its needy population. It is also underlined by WHO that a country can meet its blood requirements if just 1 percent of its eligible population donates. According to report by Office of the Registrar General & Census Commissioner, India Census, approximately 50 percent of India's population is in the age group 18-65 years which is the eligible age group for blood donation yet India fell short of 1.9 million units of blood in the year 2017. Hence, it becomes essential for India as a nation to understand the factors that dissuade people from donating blood. With proper preparation, potential donors can be identified and registered with blood donation banks.

The precise aim of this piece of study is to search for realistic and convincing features in the youths' data that could be valuable for envisaging the probability of his/her becoming a blood donor. An effort has been made to categorize candidates into donors or non-donors class on the basis of their characteristics related to blood donation. This is the first time that real datasets related to students' views, sentiments and myths towards blood donations has been collected from students of a state university of India.

The organization of the paper is mentioned here: Literature review is described in Background section. Research methodology section explains about Data Collection, Data Pre-processing, Data Analysis, Feature Extraction, Machine Learning Algorithms used for Study, Model Evaluation and Prediction, Feature Ranking. At the end, conclusions are mentioned.

## BACKGROUND

Various data mining techniques have been used extensively by researchers for classification, prediction, clustering, finding association and summarization tasks in the healthcare field. One of the unsupervised data mining techniques named k-means clustering, has been used to categorize the blood donors based on the gender, age, weight and blood group. The authors, Ramachandran *et al.* (2011), have used the datasets from Indian Red Cross Society Blood Bank. A system has been developed by ChanLee and Cheng (2011) that uses classification and clustering algorithms to determine the variations in blood donation behaviour amongst the present donors and envisage their intents towards donation so as to understand various matters and to increase the voluntary blood donation frequency. The authors have applied clustering technique to create four groups and have found that the best accuracy is 0.783.

In order to understand the awareness and attitude of students of Semnan university of medical sciences, a descriptive analytical approach has been used by Majdabadi *et al.* (2018). It was found that a large number of students are not aware of blood donation and possess a negative attitude towards blood donation.

In order to help the humanity and save precious lives, a web-based system for maintaining records of blood donors has been created by Khan *et al.* (2009). The system registers the donors and keeps their record that has details of blood donors' blood groups, address for communication, and status of blood donation. This web-enabled system acts as an interface between donors and receptors. Similar web enabled systems have been developed and deployed by Arif *et al.* (2012) and Guangpeng *et al.* (2009). With the wide spread usage of mobile communication technologies, a few notification based systems have also been deployed by Singh *et al.* (2007), Rahman *et al.* (2011), Samsudin *et al.* (2011) and Islam *et al.* (2013).

In a study by Hamouda *et al.* (2012), automatic red blood cell has been recognized and counted using image processing. Decision tree has been used to classify Red Blood Cells that has classified the data with an accuracy of 97%. Real datasets collected from an Electronic Data Processing wing of a blood bank has been classified using J48 algorithm by Sharma and Gupta (2012) that can facilitate the blood bank in-charge to make suitable decisions quicker and more accurate. There is a study by Mostafa (2009) where Intelligent data modelling techniques have been used in Egypt to examine the impact of demographic, perceptive and psychological factors on blood donations. The author has observed that there are five factors that are important for understanding blood donors' behaviour, *viz.*



Altruistic values, knowledge of blood donation, intent to donate blood, perceived risks of donation of blood, and attitude towards blood donation. A framework for the predictors for behaviour of established Australian blood donors has been determined by Masser *et al.* (2009).

Rajput *et al.* (2009) have stated that it is a great challenge to utilize data mining algorithms in the fields of healthcare and medicine. A report by Government of India (2007) has suggested that voluntary non-remunerated regular blood donations are the safest. The strategy of Indian government focuses on motivating non-remunerated blood donors and it emphasizes to maintain good epidemiological data on the occurrence of infectious markers in the general population. One of the main hindrances for blood donations are risks associated with the process as explained by Tscheulin and Lindenmeier (2005) that mainly includes fear of infection.

Qualitative studies have been used by Ferguson and Chandler (2005) to express that blood-donors depict their behaviour using Trans Theoretical Model. Schlumpf *et al.* (2007) have done extensive study based on a questionnaire filled in by approximately 8000 active donors. The possibility of return of a current donor within next 12 months has been explored using logistic regression.

The prediction of blood donor using age and blood group has been done by Sharma and Gupta (2012). The authors have made use of WEKA tool for data mining. A data mining system based on clustering and classification has been developed by Chan-Lee and Cheng (2011) [7.I] in order to understand the behaviour of blood donors.

From all these studies, it is clear that it is very important to remove myths by educating people and also to identify donors so that the blood banks and other voluntary organizations chalk out a strategy for organizing blood donation camps. By applying classification technique, the potential donors can be identified and the important factors that dissuade eligible donors from donating blood can be extracted. This research work is based on data of an Indian university's students so as to understand their knowledge, attitude and psychology towards blood donations like their fears, myths, risks while becoming a blood donor. This paper uses machine learning algorithms viz. k-nearest neighbour and logistic regression to classify potential donors as donor or non-donor. It also makes use of feature extraction to find and rank important features that play a significant role for a person to become blood donor. The ultimate objective is to motivate such eligible people to donate blood regularly so that many human lives can be saved.

## RESEARCH METHODOLOGY

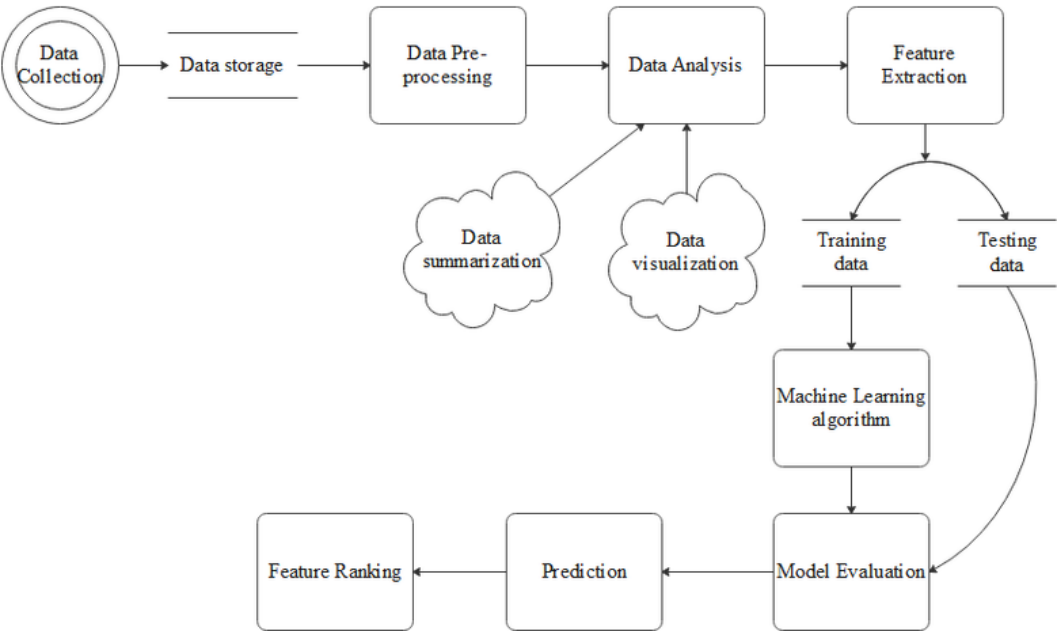
The population used for the research belongs to Generation Z (born between 1995 and 2015). These are the students of Undergraduate programmes of a Delhi state university, India. The system framework showing all steps of research in order to perform predictive analytics of blood donors, is shown in Figure 1.

### Online Collection of Data

Data has been collected by using questionnaire developed in google forms by students of an undergraduate programme of a Delhi state university in order to complete their major project. Approximately 500 students of ten different colleges, pursuing undergraduate programmes have been surveyed and responses have been gathered. Convenience sampling technique has been used and hence the selection of the participants was non-random and voluntary.

The questionnaire is based on personal attributes, intention towards blood donation, myths related to blood donation, risks associated with blood donation and perceived belief of probable donors before taking decision on blood donation. There is a total of 20 questions that includes one question describing the class of blood donor (i.e. whether the person is willing to become a blood donor or not). The response to nineteen questions is on Likert scale in which the respondents were asked to select the choice that suited them the most. The choices are Definitely yes, Probably yes, Maybe

Figure 1. System framework for research



yes, Probably no and Definitely no. The complete description of the questionnaire is mentioned in Table 10, in the Appendix.

Data Pre-Processing

Out of 20 questions, nineteen have responses on Likert scale. One-hot encoding which is a dummification technique, has been used on these nineteen features. This encoding is basically the representation of categorical variables as binary vectors. These categorical values are first mapped to integer values. Each integer value is then represented as a binary vector that is all 0s (except the index of the integer which is marked as 1). This transformation is required so as to prepare datasets for feeding to an appropriate classification algorithm in Python. The output attribute is “Willingness to become donor” that can take value viz. Yes / No. “Yes” has been transformed to “1” and “No” has been converted to “0”.

Data Analysis

Data Summarization

A total of 448 participants responded and their frequency distribution on basis of willingness to donate blood, blood groups and religion are shown in Table 1, Table 2 and Table 3 respectively.

Table 1. Distribution of willingness to donate blood as reported by participants

Willingness to donate blood	Count	Percentage
Yes	258	57.58%
No	190	42.42%

**Table 2. Distribution of blood group as reported by participants**

Blood group	Count	Percentage
A+	82	18.30%
A-	14	3.13%
AB+	57	12.72%
AB-	9	2.01%
B+	186	41.52%
B-	15	3.35%
O+	76	16.96%
O-	9	2.01%

**Table 3. Distribution of religion as reported by participants**

Religion	Count	Percentage
Buddhism	11	2.46%
Christian	10	2.23%
Hindu	346	77.23%
Jainism	24	5.36%
Muslim	19	4.24%
Sikh	34	7.59%
Other	4	0.89%

### Data Visualization

Various visualization and numerical calculations libraries of Python have been used to understand the attitude of respondents towards blood donations. These libraries are seaborn and matplotlib.pyplot for generating barplots and numpy for numerical calculations like grouping the participants on the basis of their response on likert scale. Bar plot showing percentage of respondents in each of the five options to a particular question, has been generated. There was a total of 17 such questions to understand the characteristics of participants and hence 17 such graphs are generated as shown in Table 4.

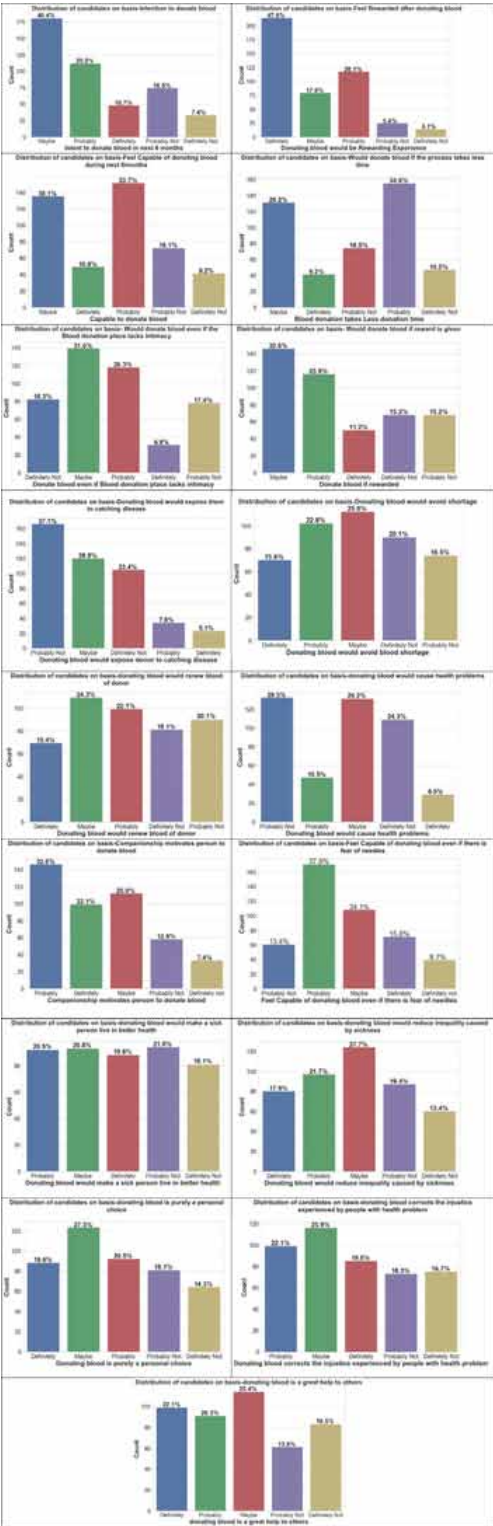
### Feature Extraction

The data visualization bar plots have been interpreted so as to select only those features for input to machine learning algorithm for classification. A few of the bar plots do not show much of variation and the distribution of respondents in each of the five categories on likert scale is of similar nature, so these features were removed before applying machine learning algorithms. The removed features are donating blood is purely a personal choice, donating blood would renew blood of donor, donating blood would avoid blood shortage.

### Machine Learning Algorithms Used for Study

Two popular machine learning algorithms viz. a lazy learner classifier (K-Nearest Neighbor) and logistic regression have been used on Spyder which is a powerful scientific environment written in Python.

Table 4. Distribution of participants in each of the five options of Likert scale, as per their response to each of the 19 questions



K-Nearest Neighbor (K-NN) is an algorithm for classification which memorizes the training data first and when presented with a testing record, it looks for similarity to the memorized training records. Whichever training record is most similar to the test case, that class is assigned to the test tuple. As explained by authors in (Peng et al., 2009) the benefit in using a lazy learning algorithm is that there is local approximation of target function for each query posted to the classifier. This leads to solving many queries in an organized and easy manner. K-NN classifier applies an incremental approach wherein the input comprises a set of attribute-value pairs, as described by Witten Eibe (2011). There is one attribute that corresponds to the class of tuple and other attributes are used as predictors.

Logistic Regression is a very popular machine learning technique based on statistics. It is a type of regression analysis method to apply when the dependent variable is dichotomous (binary). The logistic regression is a predictive analysis technique that uses a logistic function. As explained by Hosmer and Lemeshow (2000), it is used to describe the relationship between one dependent binary variable and one or more ratio-scaled, nominal, interval or ordinal independent variables.

For experimentation, K-NN and Logistic regression have been used for classifying the participants as blood donor or non-donor. The responses have been split into two parts, viz. training and testing. The training of the machine has been done with 70% of records and rest 30% have been used for testing.

## Model Evaluation and Prediction

The machine learning algorithm for classification would predict the output class of a student as either Blood donor (Positive class) or Non-donor (Negative class). There are only four categories, given below, that any student X could end up with:

- True positive (TP): Prediction is Donor and X is actually a Donor.
- True negative (TN): Prediction is Non-Donor and X is actually a Non-Donor
- False positive (FP): Prediction is Donor but X is actually a Non-Donor, so it is a false alarm.
- False negative (FN): Prediction is Non-Donor but X is actually a donor, again a wrong prediction.

These four cases in confusion matrix are shown in Table 5.

Using the confusion matrix, a number of performance metrics have been calculated in Python. These metrics are explained below.

### Accuracy

It is the ratio of the correctly labelled class to the entire collection of classes:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

### Precision

Precision is the ratio of the correctly predicted positive labelled records by the algorithm to all positive labelled records including wrongly labelled also:

Table 5. Confusion matrix

		Predicted class	
		Non-donor	Donor
Actual class	Non-donor	TN	FP
	Donor	FN	TP

$$Precision = TP / (TP + FP) \quad (2)$$

### **Recall (Sensitivity)**

Recall is the ratio of the correctly predicted positive labelled records by the algorithm to all who are actually positive in reality:

$$Recall = TP / (TP + FN) \quad (3)$$

### **F1-Score (F-Measure)**

F1 Score takes into account both precision and recall. It is the harmonic mean of the precision and recall. It is a good indicator of performance of classifier when there is uneven class distribution:

$$F1Score = 2 * (Recall * Precision) / (Recall + Precision) \quad (4)$$

### **Specificity**

Specificity is the ratio of the correctly predicted negative labelled records by the algorithm to all who are actually negative in reality:

$$Specificity = TN / (TN + FP) \quad (5)$$

### **K-NN Algorithm**

The value of k has been varied from 1 to 10 in order to find the maximum value of correctly classified records. It is found that the best classification accuracy is when k=8. The corresponding confusion matrix is shown in Table 6. Using this confusion matrix, the calculation of mentioned performance metrics has been done and is shown in Table 8.

### **Logistic Regression**

By applying this machine learning algorithm, the predicted class vs. actual class data has been shown in confusion matrix in Table 7. Using this confusion matrix, the calculation of mentioned performance metrics has been done and is shown in Table 8. As it is evident from Table 8, the K-NN algorithm with k=8 has outperformed the logistic regression in all the performance metrics.

### **Feature Ranking**

The task of determining the important features (independent) that are greatly affecting the decision of a student to be a blood donor has also been done. For this purpose, Recursive Feature Elimination

**Table 6. Confusion matrix for K-NN algorithm**

		Predicted class	
		Non-donor	Donor
Actual class	Non-donor	TN = 33	FP = 24
	Donor	FN = 16	TP = 62

Table 7. Confusion matrix for Logistic Regression algorithm

		Predicted class	
		Non-donor	Donor
Actual class	Non-donor	TN = 31	FP = 26
	Donor	FN = 17	TP = 61

Table 8. Comparison of performance measures for K-NN and Logistic Regression Algorithms

Performance measure	K-NN algorithm	Logistic Regression
Accuracy	0.7037	0.6815
Precision	0.7209	0.7011
Recall (Sensitivity)	0.7949	0.7821
F1-score (F-Measure)	0.7561	0.7394
Specificity	0.5789	0.5439

(RFE) method that generates feature importance ranking, has been used. RFE is a feature selection approach that gives the ranking of the features according to their importance for determination of the dependant variable. It uses the model accuracy to find which attributes (and combination of attributes) contribute the most to predicting the dependent attribute. There are many benefits of using RFE viz. reduction of overfitting, reduction of training time and improvement in accuracy of the model. The scikit-learn Python library provides this method. RFE model is created by using logistic regression classifier as a base model. This model has listed important features as per their ranks as shown in Table 9.

## CONCLUSION

In this paper, we have experimented with real datasets of students of Undergraduate programmes of an Indian university. A total of seventeen questions related to students' views, awareness, sentiments and myths pertaining to blood donations were asked using online questionnaire. Data visualization technique has been used to remove those features that are not contributing much towards decision of a participant to donate blood. Following this, two machine learning algorithms viz. K-Nearest Neighbor and Logistic Regression, for classification have been used to label students as donors

Table 9. Names of features as per their ranks that are main deciding factors for a person to donate blood

Rank	Feature Name
First	Would donate blood if the process of donation takes less time.
	Would donate blood as donating blood would be a rewarding experience.
	Would donate blood if accompanied by friend, family member or colleague.
	Would not like to donate blood if the place lacks intimacy.
Second	Would donate blood if it reduces inequality caused by sickness.
Third	Would donate blood if blood donation by me permits a sick person live with a better health.

or non-donors. K-NN algorithm was executed by varying value of  $k$  from 1 to 10 and the best classification accuracy was obtained with value of  $k$  equal to 8. After performing comparison of various performance metrics, it was found that the K-NN classifier has demonstrated convincing results with an Accuracy of 0.7027, Precision equal to 0.7209, Sensitivity value 0.7949, F1-score equal to 0.7561 and Specificity value 0.5789.

This study has provided the ability to identify important factors that influence the decision of youth to donate blood. These factors are time taken by blood donation process, companionship of a friend or family member while donating blood, availability of intimate place for blood donation and feeling of being rewarded. With the knowledge of these important determinants, the blood donation services can come up with newer and more efficient strategies that would increase the number of donors. This identification of probable donors would help blood banks and voluntary organizations plan in advance for the organization of blood donation camps. Also, the participants predicted as Non-donors can be motivated and the factors that restrict them from donation can be worked upon. Hence the significant gap between demand and availability of blood in India can be reduced by better management and collection of blood.



## REFERENCES

- Abolghasemi, H., Maghsudlu, M., Kafi-Abad, S. A., & Cheraghali, A. (2009). Introduction to Iranian blood transfusion organization and blood safety in Iran. *Iranian Journal of Public Health*, 38(1), 82–87.
- Agrawal, A., Tiwari, A. K., Ahuja, A., & Kalra, R. (2013). Knowledge, attitude and practices of people towards voluntary blood donation in Uttarakhand. *Asian Journal of Transfusion Science*, 7(1), 59–62. doi:10.4103/0973-6247.106740 PMID:23559768
- Arif, M., Sreevas, S., Nafseer, K., & Rahul, R. (2012). Automated online Blood bank database. *Annual IEEE India Conference*.
- Benedict, N., Usimenahon, A., Alexander, N. I., & Isi, A. (2012). Knowledge, attitude and practice of voluntary blood donation among physicians in a tertiary health facility of a developing country. *International Journal of Blood Transfusion and Immunohematology*, 2, 4–10. doi:10.5348/ijbti-2012-7-OA-2
- ChanLee, W., & Cheng, B.W. (2011). An Intelligent system for improving performance of blood donation. *Journal of Quality*, 18(2), 173–178.
- Ferguson, E., & Chandler, S. (2005). A stage model of blood donor behaviour: Assessing volunteer behaviour. *Journal of Health Psychology*, 10(3), 359–372. doi:10.1177/1359105305051423 PMID:15857868
- Government of India. (2007). *Voluntary blood donation programme*. [http://www.nacoonline.org/upload/Final%20Publications/Blood%20Safety/voluntary%20blood%20do nation.pdf](http://www.nacoonline.org/upload/Final%20Publications/Blood%20Safety/voluntary%20blood%20do%20nation.pdf)
- Guangpeng, L., Zhongwen, G., Song, X., & Wenli, P. (2009). Web-based real-time monitoring system on cold chain of blood. *IEEE Instrumentation and Measurement Technology Conference*. doi:10.1109/IMTC.2009.5168655
- Hamouda, A., Khedr, A. Y., & Ramadan, R. A. (2012). Automated Red Blood Cells counting. *International Journal of Computational Science*, 1(2), 13–16.
- Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). John Wiley & Sons, Inc.
- Islam, A., Ahmed, N., Hasan, K., & Jubayer, M. (2013). mHealth: Blood donation service in Bangladesh. *International Conference on Informatics, Electronics and Vision (ICIEV)*. doi:10.1109/ICIEV.2013.6572594
- Khan, A.R. (2009). Web based Information System for Blood Donation. *International Journal of Digital Content Technology and Its Applications*, 3(2), 137–142.
- Majdabadi, H. A., Kahouei, M., Taslimi, S., & Langari, M. (2018). Awareness of and attitude towards blood donation in students at the Semnan University of Medical Sciences. *Journal of Electronic Physician*, 10(5), 6821–6828. doi:10.19082/68 PMID:29997767
- Masser, B. M., White, K. M., Hyde, M. K., Terry, D. J., & Robinson, N. G. (2009). Predicting blood donation intentions and behaviour among Australian blood donors: Testing an extended theory of planned behaviour model. *Transfusion*, 49(2), 320–329. doi:10.1111/j.1537-2995.2008.01981.x PMID:19040598
- Mostafa, M. M. (2009). Profiling blood donors in Egypt: A neural network analysis. *Expert Systems with Applications*, 36(3), 5031–5038. doi:10.1016/j.eswa.2008.06.048
- Nigatu, A., & Demissie, D. B. (2015). Knowledge, Attitude and Practice on Voluntary Blood Donation and Associated Factors among Ambo University Regular Students, Ambo Town, Ethiopia. *Journal of Community Medicine & Health Education*.
- Peng, Y., Jianyong, Z., & Yumhong, X. (2009). Lazy learner text categorization algorithm based on embedded feature selection. *Journal of Systems Engineering and Electronics*, 20(3), 651–659.
- Rahman, M. S., Akter, K. A., Hossain, S., Basak, A., & Ahmed, S. I. (2011). Smart blood query: a novel mobile phone based privacy-aware blood donor recruitment and management system for developing regions. *IEEE Workshops of International Conference*. doi:10.1109/WAINA.2011.115
- Rajput, A., Aharwal, R. P., Chandel, N., Solanki, D.S., & Soni, R. (2009). Approaches of Classifications to Policy of Analysis of Medical Data. *International Journal of Computer Science and Network Security*, 9(11), 1–9.

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2 • July-December 2021

Ramachandran, P., Girija, N., & Bhuvaneswari, T. (2011). Classifying blood donors using data mining techniques. *International Journal of Computer Science and Engineering Technology*, 1(1), 10–13.

Samsudin, N., Khalid, S. A., Yusoff, A. M., Ihkasan, M., & Senin, Z. (2011). Procedure automation with immediate user notification: A case study. *IEEE Symposium on Business, Engineering and Industrial Applications (ISBEIA)*. doi:10.1109/ISBEIA.2011.6088816

Schlumpf, K. S., Glynn, S. A., Schreiber, G. B., Wright, D. J., & Randolph, S. W. (2007). Factors influencing donor return. *Transfusion*, 48, 264–272. PMID:18005325

Sharma, A. & Gupta, P.C. (2012). Predicting the number of blood donors through their age and blood group by using data mining tool. *International Journal of Communication and Computer Technologies*, 1(2).

Singh, R., Bhargava, P., & Kain, S. (2007). Smart Phones to the Rescue: The Virtual Blood Bank Project. *IEEE Pervasive Computing*, 6(4), 86–89. doi:10.1109/MPRV.2007.87

Tscheulin, D., & Lindenmeier, J. (2005). The willingness to donate blood: An empirical analysis of socio-demographic and motivation-related determinants. *Health Services Management Research*, 18(3), 165–174. doi:10.1258/0951484054572547 PMID:16102245

Verma, S., Sharma, R. K., Sharma, M., & Pugazhendi, S. (2016). Voluntary Blood Donation: Attitude and Practice among Indian Adults. *Journal of Community Medicine & Health Education*, 6(3). Advance online publication. doi:10.4172/2161-0711.1000436

Witten, I. H., & Eibe, F. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). The Morgan Kaufmann Series in Data Management Systems.

## APPENDIX: DESCRIPTION OF QUESTIONNAIRE

Table 10. Description of the questionnaire

S.N e	Variable	Description	Collected Data/ Response
1	BGroup	Blood Group	<input type="radio"/> A+ <input type="radio"/> A- <input type="radio"/> AB+ <input type="radio"/> AB- <input type="radio"/> B+ <input type="radio"/> B- <input type="radio"/> O+ <input type="radio"/> O-
2	Religion	Religion	<input type="radio"/> Buddhism <input type="radio"/> Christian <input type="radio"/> Hindu <input type="radio"/> Jainism <input type="radio"/> Muslim <input type="radio"/> Sikh <input type="radio"/> Other
3	Intend_to_give	Intend to give blood in next 6 months	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
4	Rewarding_exp	For me, donating blood during the 6 months would be a Rewarding Experience	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
5	Capability	I am eligible to donate blood during the next 6 months	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
6	Less_donation_time	If the process of donating blood took less time , I would donate blood	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
7	Intimacy_level	I would agree to donate blood even if blood donations takes place, in a place, which lacks intimacy	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
8	Given_a_reward	If I were given a reward, I would donate blood	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
9	Catch_a_disease	If I were to give blood, this would expose me to catching a disease	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
10	Avoid_shortage	If I were to donate blood, this would help avoid blood shortage	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
11	Renew_my_blood	If I were to donate blood, this would help me to renew my blood	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
12	Cause_health_problems	If I were to donate blood, this could cause me health problems	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
13	Companionship	If I were accompanied by a friend, a family member or a colleague, I would donate blood.	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
14	Trypanophobia	I would be capable of donating blood, even if I were afraid of needles or fainting.	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
15	Improved_sick_life	If blood donation by me permits a sick person live in a better health, I would donate blood.	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
16	Reduced_inequality	If blood donation by me reduces inequality caused by sickness, I would donate blood.	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
17	Personal_choice	If blood donation is purely my personal choice, I would donate blood.	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
18	Injustice_by_health	If blood donation by me corrects the injustice experienced by people with a health problem that necessitates a blood transfusion, I would donate blood.	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
19	Help_people	If blood donation by me helps my fellow men/women I would donate blood.	<input type="radio"/> Definitely <input type="radio"/> Probably <input type="radio"/> Maybe <input type="radio"/> Probably not <input type="radio"/> Definitely not
20	Blood_Donor	Willing to become a blood donor	<input type="radio"/> Yes <input type="radio"/> No

*Kavita Pabreja is working as Associate Professor in the Department of Computer Applications at Maharaja Surajmal Institute, an affiliate of GGS Indraprastha University, New Delhi, India. She received her Ph.D. in Computer Science from Birla Institute of Technology & Science, Pilani. She holds more than 25 years of experience which includes teaching experience of 20 years and industry experience of over 5 years with Indian as well as USA MNC. She has authored four books, two with International publishers - "Application of Artificial Intelligence tools – Impact on Weather Prediction", "Object-Oriented Programming Using C++", and two with National publishers- "Learning Visual Basic.Net Programming", "Front End Design Tool VB.NET". She has contributed over sixty papers in International Journals / Book/ International conferences of repute. She is a recipient of a Best Research Paper Award from EMC data storage systems and Amity for All India Competition in Data Research; another Best Research Paper award for developing An Android Application to illustrate the Perspective of Community on Blood Donation at National Conference organized by IITM, affiliate of GGSIP University, India.*

*Akanksha Bhasin is currently a Post Graduate student, pursuing Masters of Computer Applications in Software Engineering at University School of Information, Communication and Technology, GGSIP University. She completed her graduation in the field of Computer Applications from Maharaja Surajmal Institute, an affiliate of GGS Indraprastha University, Delhi. She received Academic Excellence Award from MSI as well as Exemplary Performance Award from GGSIP University for the same for her outstanding academic achievements. Moreover, being a data science enthusiast, she published numerous research papers and has been conferred with Best Research Paper award for developing "BloodMate- An Android Application To Illustrate the Perspective of Community on Blood Donation" at National Conference organized by IITM, an affiliate of GGSIP University, Delhi.*

# Comprehensive Analysis of State-of-the-Art CAD Tools and Techniques for Chronic Kidney Disease (CKD)

Mynapati Lakshmi Prasudha, VNR Vignana Jyothi Institute of Engineering and Technology, India

Rakesh Kasumolla, VNR Vignana Jyothi Institute of Engineering and Technology, India

Deepak Sukheja, VNR Vignana Jyothi Institute of Engineering and Technology, India

## ABSTRACT

In the last one decade, AI/ML/DL has been considered a core research area in healthcare. As we know that the kidneys are important internal body organs that help in regulation of the fluid within the body such that they relieve the body from the existence of waste. Disease is difficult to detect early on by normal clinical process. Many researchers have focused their work to identify kidney disease or classify kidney disease using computational technology because the mortality rate is very high in kidney patients. The primary focus of this paper is to review the current research work based on computational advancement in the area of kidney disease and also identify the gaps or future scope to improve the process of classifying kidney disease at earlier stage.

## KEYWORDS

CADs, Convolutional Neural Network (CNN), DL/AI/ML in Healthcare, Kidney Disease, Medical Image Processing

## 1. INTRODUCTION OF CKD

Initial intervention generally reduces severe disease progress. A report published by Elsevier in February 2020 it shown in figure 1, is indicating the worldwide Growth of Kidney Disease problem is alarming. Near about 8 million peoples are affected by various kinds of kidney related diseases most of kidney disease categorized as Chronic Kidney Disease (CKD) or Acute Kidney Injury (AKI). Chronic Kidney Disease are progressive and non-recoverable in most of cases and Acute kidney injury are 100% recoverable. But it is also observed that during the treatment of Acute kidney injury around 30% to 40% patients also started to suffer the Chronic Kidney Disease (CKD), which is very serious and danger situation for the patients.

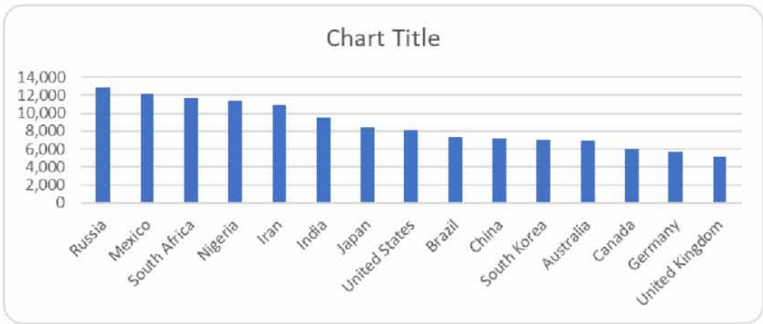
Prevalence rates for chronic kidney disease in select countries worldwide in 2017 (per 100,000 population)

CKD is a condition where the gradual loss is being observed in the functionality of the kidneys disallows the purification of the blood. CKD is one of the clinical problems considering the process of dealing with the disease at its final phase and from top to toe likelihood of expiry (Bikbov & Vos, 2020; Bulletin of the World Health Organization, 2018). Due to kidney disease, according to the WHO (World Health Organization) incidents and deaths reported in 2018 in United States only shown in Figures 2 and 3, respectively.

DOI: 10.4018/IJBDAH.287605

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Figure 1. Chronic kidney disease prevalence rates for select countries worldwide 2017



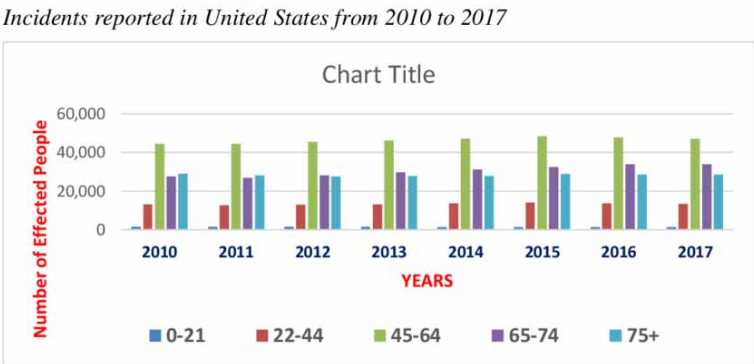
*Chronic kidney disease prevalence rates for select countries worldwide 2017*  
*Prevalence rates for chronic kidney disease in select countries worldwide in 2017 (per 100,000 population)*

According to a survey by (National Kidney Foundation, 2002), 60% cases are addressed in undeveloped and developing countries. In Bangladesh around 18% of the peoples are affected with kidney disease and among them most of the patients are affected with DM (Diabetes Mellitus Type 1) (39.02%) and DM-2 (Diabetes Mellitus Type 2) (41.46%). In India, the rise in mortality rate due to chronic diseases has been observed as 3.78 million in 1990 (40.4% of total deaths) to 7.63 million in 2020.

1.1 Risk Classification of CKD

It has been observed that Antivirals, Hyperventilation or Medications and habdomyolysis can also led to kidney injury. Therefore, as per WHO reports many Acute kidney injury cases has been reported during the treatment of COVID-19 also.

Figure 2. Incidents reported in United States from 2010 to 2017



Several articles have been published (Assmann, Cullen, & Schulte, 2002; Initiative KDOQI. K/doqi clinical practice guidelines on hyper-tension and antihypertensive agents in chronic kidney disease, 2004; Zandi-Nejad, Luyckx, & Brenner, 2006; Hippisley-Cox et al., 2008; Deng et al., 2017) about CKD risks & its analysis and authors have observed in their research, hypertension and diabetes are the most common cause of the Chronic Kidney Disease. Apart of hypertension and diabetes, numbers of factors increase the risk of affecting kidneys such as obesity, smoking, old age, high blood pressure (hypertension), certain inherited diseases, family history of cancer and so on. In medical domain, to identify the kidney disease, nephrologist (doctors) suggests the different medical reports based on patient's situation. These reports are urine(urinalysis) key parameters are (RBCs - £2 RBCs/hpf, WBCs - £2-5 WBCs/hpf), blood tests, imaging tests such as CT scans or MRIs, comprehensive metabolic panel, urine culture, complete blood count, liver or renal panel, Antibiotic Susceptibility Testing and kidney biopsy. Risk Classification of CKD and Evaluation Plans are mentioned in Table 1.

As we know initial intervention generally reduces severe disease progression. In the view of the technological advancements for diagnosing the KD (Kidney Disease) several automated tools can help to identify and classify Kidney Disease Because of in healthcare various tools has been implemented to classify the multiple diseases using AI, Deep Learning and Machine Learning i.e advanced computational technologies can play an important role to predict Kidney Disease at earlier stages.

## 2. COMPUTER-AIDED DIAGNOSIS TOOLS AND TECHNIQUE

Upon reviewing, it is observed that Machine learning, the subset of artificial intelligence could develop relationships using the data where defining them priorly would not be necessary (Deng et al., 2017; Bhaskar & Manikandan, 2019). Machine learning is classified into three types namely **Supervised Learning**, whose objective is to learn a mapping from inputs  $x$  to outputs  $y$ , input-output pairs  $D = \{(X_i, Y_i)\}_{i=1}^n$ . Here  $D$  denotes a training set, and  $N$  training examples. **Unsupervised Learning** is another type in which only inputs are provided  $D = \{(X_i)\}_{i=1}^n$  and to find “interesting patterns” from the data. **Reinforcement learning** is another type that plays a very crucial role in predicting the behavior of the data.

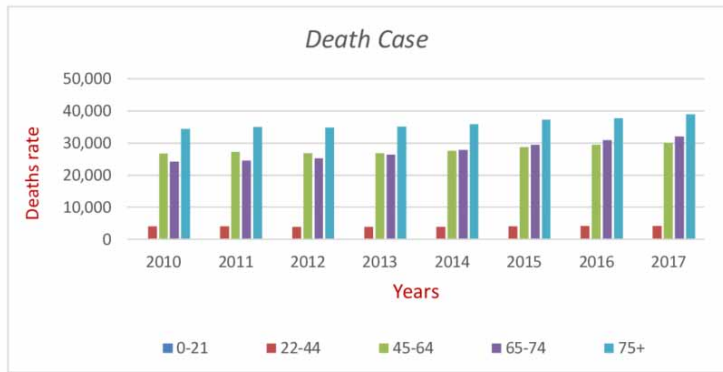
Deep learning (DL), one of the applications of artificial intelligence, relies on algorithms to process the data and recreation for emerging concepts. It allows computational models that can be composed of multiple processing layers using neural networks that include variety of techniques. The most recognized algorithm, convolutional neural network (CNN) is a part and parcel of Deep Learning models that has become an indispensable process in tasks comprising image detection. In (Khalifa, Taha, Ezzat Ali, Slowik, & Hassanien, 2020), It states that there are three key methods that successfully work CNNs for image classification: a) training the “CNN from scratch” b) using “off-the-shelf CNN” features extraction as corresponding information channels disease identification and 3) performing fine-tuning of target images. Architecture for the classification of images is as shown in Figure- 3. Working procedure and functionality of CNN are also well defined in (Khalifa, Taha, Ezzat Ali, Slowik, & Hassanien, 2020).

CNNs consist of convolutional layers which are characterized by an input map  $I$ , a bank of filters  $K$  and biases  $b$ . As mentioned by Jekfine in their article, If input as type of image with height  $H$ , width  $W$  and  $C$  (red, blue, and green) such that  $\hat{I} \in \mathbb{R}^{H \times W \times C}$ . Subsequently for a bank of  $D$  filters we have  $\hat{K} \in \mathbb{R}^{k_1 \times k_2 \times C \times D}$  and biases  $\hat{b} \in \mathbb{R}^D$ , one for each filter. Neural Network is a super set of all types of deep learning approaches. The major benefits of neural networks over conventional programming are problems solving ability. Therefore, Neural network technology is seen as cutting-edge today. Neural networks efficiently handle problems like Prediction and pattern recognition. But neural network also has some limitations that it cannot apply neural network or deep learning technique everywhere because of huge complication. To get the higher accuracy of result from deep learning approaches, the size of data set should be very large i.e deep learning approaches are not suggested for small or

**Table 1. Risk classification of CDK and evaluation Plans**

S. No	Stages of CIO	Glomerular filtration rate	Action Plans
1	Kidney injury with normal GFR	90 and above	Analysis of comorbid situations, phylogenies, decrease in risk
2	Failure of Kidney with mild decrease	60 to 89	Estimate disease evolution
3	Moderate reduction	30 to 59	The Valuation and dealing of ailment problems
4	Severe lessening	15 to 29	Research in additional treatments like dialysis and replacement of the kidney for Renal failures.
5	Kidney fiascos	Less than 15	Kidney standby therapy

**Figure 3. Death case reported in United Ststes from 2010 to 2017**



*Death case reported in United Ststes from 2010 to 2017*

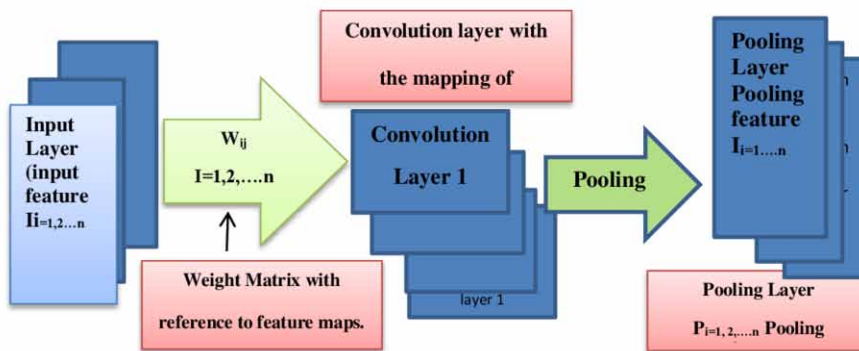
average size datasets owing to the higher levels of training requirements for neural networks. After the major success in 2013, KAIST University in South Korea, Deep Learning experts trained their deep learning model with chest x-rays and mammography images for the detection of lung and breast cancer with an accuracy of 97%. Several researchers have initiated their research in the direction of predicting the kidney disease at earlier stages using AI, Deep Learning, Machine Learning and fuzzy computation.

### 3. COMPUTATIONAL TECHNOLOGY GROWTH TOWARDS CKD DIAGNOSIS

Taking advantage of the information from various types of data for the betterment of the process of diagnosing Deng, Su-Ping, et al. in Deng et al., 2017 mentions the utilization of clinical and pathological features evaluation of Genomic alterations, DNA methylation profiles, RNA and proteomic signatures in KIRC. Gene expression profiles, DNA methylation expression and clinical data are downloaded from TCGA data portal. To predict the cancer stage primarily, network constructed from gene expression data, network constructed from DNA methylation data, secondarily, order to investigate the potential of using networks as a diagnostic tool fused networks (network constructed



Figure 4. CNN

*An architecture and the process of training of CNN (convolutional neural network)*

from DNA and network constructed from gene expression) from two data types: gene expression and DNA methylation. In this paper it is claimed that they have achieved a good accuracy in the prediction of the KIRC cancer stage via the fused network. The data taken by the authors is not sufficient, so the authors has used the simulated data which cannot be consider during the experiment/ evaluation of critical disease.

In (Zhang et al., 2019) Hui Zhang, et al., it is suggested that morphological cascade CNNs are based on multi-intersection over union (IOU) threshold in their works. For the identification of minor injuries (1-5 mm), 2 convolution layers in morphology, adapted Feature Pyramid Networks (FPNs) in faster RCNN and combined 4 IOU threshold cascade RCNNs are being implemented. For experimentation, medical CT scan image data set, published by NIH (National Institutes of Health) has been utilized. Experimentation has a requirement of Hardware with Intel Core i5, 2.7GHz CPU, 8GB RAM that supports Ubuntu, a NVIDIA GTX 1080 video processing card and software's that includes Cascade RCNN deployed in pytorch 1.0 framework, and Faster RCNN used for tensorflow - GPU 1.8. And both python3.5, cuda 9.0 and cudnn 7.1.4. MATLAB based frameworks are utilized. In the work the authors have mentioned about the damage done to the kidney through CNN. As we know that CNN or DL requires large input set but, in this experiment only 985 kidney images were used, which is not enough.

For the prediction of CKD Himanshu Kriplani et al., (Kriplani, Patel, & Roy, 2019) proposed Deep Neural Network. In 2015 UCI Machine Learning Repository named Chronic Kidney Disease data set is used for proposed Deep Neural Network. The data set comprises of 400 instances, 25 attributes in which 11 are numeric and 14 are nominal. All instances were classified into two categories; 105 has label CKD and 119 has label NCKD (Non CKD). Categories data set is split into two parts 60% and 40%, 60% used for training purpose and 40% for testing. With the help of the confusion matrix, the overall performance of the model with an accuracy of 97% has been measured. The authors mentioned that 18 parameters used during the creation of model but they did not mentioned the details of any parameters.

Navaneeth Bhaskar and Suchetha M et al., in (Bhaskar & Manikandan, 2019) proposed a new sensing technique for the automatic detection of CKD. The technique is based on a one-dimensional CNN algorithm and SVM classifier. To test the technique, authors collected 102 samples, including 40 healthy volunteers and 62 individuals with CKD. Samples are dropped via input opening of the chamber for the process of testing which relies on the amount of ammonia gas produced, changes in

the electrical conductivity of the gas sensor. Deep learning CNN–SVM algorithm is used for automatic computation and classification of the features from the output signal of the sensor.

In (Khalifa, Taha, Ezzat Ali, Slowik, & Hassanien, 2020), **Nour Eldeen M. Khalifa et al** suggested a novel optimized approach for the process of classification of 5 various categories of renal cancer, clear cell carcinoma (KIRC), breast invasive carcinoma (BRCA), lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD) and uterine corpus endometrial carcinoma (UCEC) with respect to available images in the dataset. The approach is implemented in 3 phases, preprocessing, augmentation, and deep CNN architecture. Under the preprocessing technique, the high dimensional RNA sequence is optimized to select the optimum number of features using BPSO-DT and then, converts the RNA-Seq into 2D images. The augmentation phase increases the original dataset of 2086 samples to 5 times larger, and finally developed deep CNN architecture to classify KIRC, BRCA, LUSC, LUAD and UCEC cancer. To perform the experiment, researchers used a software package (MATLAB), Intel Xeon E5-2620 processor (2 GHz), 32 GB of RAM and 12 GB Nvidia GTX Titan X GPU speciðc. To check the accuracy of proposed approach, confusion matrix model is used. Researchers claim minimum 95% accuracy to classify all the 5 types of cancer.

Ya feng Ren, HaoFei et al presents a hybrid neural network model in (Ren, Fei, Liang, Ji, & Cheng, 2019). In their works, researchers analyzed the relationship between hypertension and kidney disease. They Collected dataset of patients suffering from diseases or only hypertension. Center points of proposed hybrid neural network model are reimplementatoin of baseline systems, design the discrete model using Naïve Bayes (NB), Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT) and neural network using CNN. The accuracy varies from 64.9 to 81.2% with respect to different models and features. The Model used in the experiments are LR, NB, SVM and GBDT and feature with respect to every model are Numerical, Textual and Textual-Numerical.

In (Ravizza et al., 2019), it is said that comparison of predictive analytic algorithms using real data could achieve equivalent or enhanced accuracy which is distinguishable with clinical trial data. The accuracy provided for the identification of CKD risk Red, Roche/IBM algorithm is compared with alternative algorithms proposed by Dunkler et al. (Han, Hwang, & Lee, 2019), Vergouwe et al.(Thong, Kadoury, piche, & pal, 2018), d Keane et al.(Ravishankar, Prabhu, Vaidya, & Singhal, 2016) and Jardine et al.(Ravishankar et al., 2016). Roche/IBM algorithm performs better when compared to others. Also, in (Weng, Reps, Kai, Garibaldi, & Qureshi, 2017) Stephen F. Weng et. Al. tried to assess machine-learning algorithms and check cardiovascular risk prediction which can be improved through ML. complexity, Predictive accuracy and sensitivity of Random forest, logistic regression, gradient boosting machines, neural networks algorithms were compared for the same medical data (24,970 incident cardio-vascular events). In 2017 Geert Litjens et. Al in (Litjens et al., 2017) published an article “A survey on deep learning in medical image analysis” reviewed the major deep learning concepts pertinent to medical image analysis.

Comparison of various research techniques along with their data source and data size and critics shown in table 2.

#### 4. FINDINGS AND FUTURE DIRECTIONS

It is believing that artificial intelligent (AI) playing a significant role in the healthcare offerings. Various deep learning and machine learning models based on Computer-Aided Diagnosis techniques for CKD have been reported. Most of the proposed works are based on segmentation via CNN or ML, their effectiveness in terms of the datasets used, although the developed models are learning on smaller datasets. Well-defined large CKD patients (real-time) datasets are still needed. Comparing accuracy of previously reported model by many researchers is highly challenging, due to variances in data set and their characteristics as no direct quantitative metrics or mechanism are available for comparison. Building such common datasets would be a time consuming and a cost ineffective activity. Data and the truth tables would be considered as most significant elements for the consideration to

Table 2. 4 Year Survey on Technological Growth Toward Ckd Diagnosis

Author Name and Publication Year	Applied Computational Techniques/ algorithms	Source of Dataset	Research approach
Ozgun Cicek et al.(Olaf, 2016), 2016	Neural Network (deepl learning)	South African toad Xenopus laevis	Authors introduced concept of an end-to-end learning, that semi-automatically and fully-automatically segments a 3D volume from a sparse annotation. It offers an accurate segmentation for the highly variable structures of the Xenopus kidney. In this article proposed u-net architecture as an extension for semi- automatic and full automatic segmented architecture. Accuracy measured in terms of Intersection over Union (IoU) of their architecture. Researchers collect 248 sample of “Xenopus kidney” dataset from South African toad Xenopus laevis in the form of $132 \times 132 \times 116$ voxel images and used as an input to proposed architecture with 3 channels and output generated voxels images $44 \times 44 \times 28$ at final layer is in x, y, and z directions. Respectively. <b>Weighted softmax loss function</b> is used in training phase and researcher claimed accuracy is 0.863 in 3-fold cross validation.
Xiaoguang Lu et al.(Lu,Xu, & Liu.), 2016	CNN and FCN ((deep learning)	Sample Collection	Research presents a new framework based on CNN and FCN called dual learning architectures and fusion for organ detection. To localize the organ probabilistic graphical model in proposed architecture. A right kidney CT body scan image contains a stack of axial slices used as Inputs in CNN model. 450 samples were collected from 450 patients to train the CNN models.
Fang Lu et al., 2016	Neural Network (deep learning) and Graph cut	MICCAI-Sliver07 and 3Dircadb1	In this article, 3D deep CNN-based technique is used to identify the liver location and perform the segmentation task through 3D graph cut based segmentation refinement. Researcher used 40 CT scan images from public data set source available on MICCAI - Sliver07 and 3Dircadb1. Parameters used in the result analysis are VOE, RVD, ASD, RMSD and MSD. The proposed technique performed much faster when compared to manual segmentation system but in number of cases they fail to segment CT images. Similar kind of task (localization of Anatomical Structures in 3D Medical Images) is performed on different dataset with a different approach.
Hariharan Ravishankar et al., 2016	Neural Network (deep learning)	LOGIQ E9 scanner	This article mentions the implementation of a new hybrid deep CNN model with CaffeNet features to automatically segment and quantify abdominal shape in foetal ultrasound images. In this work, authors used Ultrasound data as an input for their model. Sample size used is 90 and source of input (Ultrasound data) was LOGIQ E9 scanner <a href="https://www.gehealthcare.in/products/ultrasound/logiq/logiq-e9-with-xdclear">https://www.gehealthcare.in/products/ultrasound/logiq/logiq-e9-with-xdclear</a> . Merits of deep learning and conventional features are also explored in this article.
Hariharan Ravishankar et al., 2017	Neural Network (deep learning)	2-D fetal ultrasound images	This article presents the extended work and proposed a hybrid approach combining traditional texture analysis methods using deep learning and HOG – GBM features. 70 samples of 2-D foetal ultrasound images used to check accuracy of proposed technique. To avoid the vanishing gradient problem, they used ReLU as the activation function.
William Thong et al., 2016	Neural Network (deepl learning)	Sample Collection	To diagnosis kidney disease, fully automatic framework presents for kidney segmentation using CNN. To predict the class membership model trained through patch-wise approach. 3 random subsets were generated from 79 sample data. Present framework is fully based on 2 D inputs and evaluation of segmentation is based on linear, Nearest Neighbour, Bilinear interpolation methods.
Su-Ping Deng et al., 2017	Network fusion method and Generalized linear model.	TCGA	Proposed a novel concept called Network fusion method. The work done is about the cancer stage prediction in the kidney. To predict the label of a new sample semi-supervised learning is used. The results have proved that Network-based LASSO Label Prediction (NLLP) method has good potential. Major problem with this work is results have generated through simulated data. Simulated data cannot be trusted for treatment in a dangerous disease like cancer.
Stephen F. Weng et al., 2017	Machine-learning Algorithms	Clinical Practice Research Data link (CPRD)	Compared different established machine-learning algorithms like random forest, logistic regression, gradient boosting machines, and neural networks to predict first cardiovascular event over 10-years. Comparison was performed on 3,78,256 CPRD dataset and findings risk prediction improved by random forest +1.7%, logistic regression +3.2%, neural networks +3.6% and by gradient boosting +3.3% . This risk prediction is based on 29 parameters.
Timothy L. Kline et al., 2017	Deep Neural Network	TEMPO study	Described a fully automated deep Neural Network approach for Fully Automated Segmentation of Polycystic Kidneys disease. Researchers simulated MR Image with multi-observer approach to predict an accurate result. 2000 DICOM image dataset is used in Deep Neural Network. Proposed approach offered effective process to measure the TKV imaging biomarker for kidneys patients.

continued on next page

# International Journal of Big Data and Analytics in Healthcare

Volume 6 • Issue 2

**Table 2. Continued**

Author Name and Publication Year	Applied Computational Techniques/ algorithms	Source of Dataset	Research approach
Eli Gibson et al., 2018	Deep Learning	TCIA and BTCV	To segment multiple organs, proposed a registration-free deep-learning-based segmentation algorithm based on dilated convolutions, dense feature stacks, batch-wise spatial dropout, up-sampling and V-network down-sampling. During the experimental study, liver, gallbladder, spleen and left kidney types of organs have been segmented. Adam optimizer is used to train the network. In this study, ninety abdominal CT images of the spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas and duodenum are used. Proposed technique is compared with deep-learning-based VoxResNet and MALF-based DEEDS +JLF pro.
Hui Zhang et al., 2019	Morphological image processing technique and Neural Network (deep learning)	Deep Lesion (PACSs)	Author proposed a Framework of kidney lesion detection, that mainly includes two steps: first acquisition of six different sizes lesions detection feature maps and second multi-IOU cascade RCNN. In the work the authors have mentioned about the damage done to the kidney through CNN. As we know that CNN or DL requires large input set but, in this experiment only 985 kidney images were used, which is not enough
Navaneeth Bhaskar et al., 2019	Machine learning technique and Neural Network (deep learning)	Sample Collection	Proposed new electronics-based sensing mechanism for the automated detection of kidney disease, and it monitors the urea levels in the saliva sample. The sensing mechanism consists of a gas sensing chamber, Arduino board and an MQ-series ammonia gas sensor. The conversion reaction is carried out inside the gas chamber. Testing is performed by dropping the sample through the input opening of the chamber for normal sample sensor produces a voltage of 0.62–0.93 V. Their results are promising with good accuracy but there is need for sample testing towards accuracy of analysis of this approach.
Yafeng Ren et al., 2019	Hybrid neural network	EHR Data of 12 Hospitals (2012-2017)	Proposed Hybrid neural model based on (BiLSTM) and Auto encoder networks for identifying kidney disease in hypertension patients. In proposed model mainly includes BiLSTM and Autoencoder. BiLSTM is used for learning and deal with textual description information and Autoencoder deal with physical indicators. Used 35,332 HER data sample to check the accuracy of implemented model. The proposed model may be used for coronary heart disease prediction in hypertension patients.
Stefan Ravizza et al., 2019	Roche/IBM algorithm	IBM Explorys database	Researchers presents a novel Roche/IBM algorithm, implemented in IBM lab. Algorithm was tested on 550,00,000 INPC huge data set. Roche/IBM algorithm is used over the real-world data and achieved the good accuracy
Seokmin Han et al., 2019	Deep Learning	Seoul National University Bundang Hospital	Proposed deep learning approach using convolutional neural network and ROI for three-layer classification to classify the renal cancer in kidney. In the article, crucial information about data set is missing.
Nour Eldeen M. Khalifa et al., 2020	Deep learning approach	Tumor gene expression dataset	Authors have proposed a common framework in his research work to classify five different types of cancer. To know the accuracy and precision of the proposed framework, the authors have used 2086 sample, in which KIRC (kidney related dataset) number is only 537, it is not enough for a deep learning techniques.
Shi Yin et al., 2020	Boundary distance regression and Pixel classification networks	US images	Proposed boundary distance regression and Pixel classification networks approach. The approach is good but has average performance and significantly better than deep learning-based pixel classification networks. To check the performance 185 sample of Kidney Pole images were used.
Ho Sun Shon et al., 2020	Deep Learning	TCGA	Proposed an end-to-end, cost-sensitive hybrid deep learning (COST-HDL) approach with a cost-sensitive loss functions. This approach is more efficient compared to traditional data mining techniques and conventional machine learning. To check the efficiency authors used Gene expression data with 1157 sample size.
Guozhen Chen et al.	Neural Network (deep learning)	miRNA genome data	Authors has proposed Adaptive hybridized Deep Convolutional Neural Network (CNN) for the early detection of Kidney disease. authors have used datasets that present at <a href="https://www.mediafire.com/">https://www.mediafire.com/</a> . The proposed AHCNN explore the consistent renal cell rating forecasts from CT (CECT) images. ACCURACY ANALYSIS is well defined in result experiments. Efficiency of the proposed model at all classification thresholds level is explained through ROC curve.

apply deep learning or machine learning methods. The inadequate datasets are a major problem to further advancement of Computer-Aided Diagnosis techniques and models In Chronic Kidney Disease (CKD). Finally, upon reviewing the various literature, it can be said that the prediction of CKD can be improved in different ways only with the development of machine learning or deep learning models. It can also be made possible through exploring the performance-complementary properties (Hybrid Model) using a combination of deep learning model, fuzzy logic and statistics techniques (lot of classical statistical methodology are present through which establish an equation or curve) for scratch medical image, pathological microscopic image and text data dataset.

## 5. CONCLUSION

Computer-Aided Diagnosis Tools and Technique for Chronic Kidney Disease have adept various domains including medical research with magnificent achievements, and an interest has emerged progressively in radiology. Although most of researcher have focused on deep learning method to classify the CKD and proposed the good model also. Most of the research is facilitating learning on smaller datasets because of the non-availability of sufficient dataset. one more interesting factor is maximum research is based on medical image data, no one have considered the pathological dataset during their research. As per doctor opinion to predict the kidney diseases at right time or earlier stage pathological report behavior of patient must be observed. Sometime “pathological based data” indicate abnormality at earlier stage but to correctly classify the abnormality medical images have been required. Similarly, sometime medical images indicate abnormality at earlier stage but to correctly classify the abnormality pathological based data have been required. As a research point of view, lot of scope is available to identify and classify kidney disease using advanced computational techniques.

Although deep learning has become a dominant method to classify the CKD and several high-profile successes of deep learning and machine learning technique have been reported in table 2, most of the research is facilitating learning on smaller datasets.

## REFERENCES

- Alonso, A., Lau, J., Jaber, B. L., Weintraub, A., & Sarnak, M. J. (2004). Prevention of radiocontrast nephropathy with N-acetylcysteine in patients with chronic kidney disease: A meta-analysis of randomized, controlled trials. *American Journal of Kidney Diseases*, 43(1), 1–9. doi:10.1053/j.ajkd.2003.09.009 PMID:14712421
- Assmann, G., Cullen, P., & Schulte, H. (2002). Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (procam) study. *Circulation*, 105(3), 310–315. doi:10.1161/hc0302.102575 PMID:11804985
- Bhaskar, N., & Manikandan, S. (2019). A deep-learning-based system for automated sensing of chronic kidney disease. *IEEE Sensors Letters*. Letters, 3(10), 1–4. 10.1109/LSENS.2019.2942145
- Bikbov, B., & Vos, T. (2020). *Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study*. Academic Press.
- Chen, G., Ding, C., Li, Y., Hu, X., Li, X., Ren, L., Ding, X., Tian, P., & Xue, W. (2020). Prediction of chronic kidney disease Using Adaptive Hybridized Deep Convolutional Neural Network on the in-ternet of medical things platform. *IEEE Access: Practical Innovations, Open Solutions*, 8, 100497–100508. doi:10.1109/ACCESS.2020.2995310
- Chen, H., Dou, Q., Yu, L., Qin, J., & Heng, P. A. (2018). VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170, 446–455. doi:10.1016/j.neuroimage.2017.04.041 PMID:28445774
- de Vos, B. D., Wolterink, J. M., de Jong, P. A., Leiner, T., Viergever, M. A., & Išgum, I. (2017). ConvNet-based localization of anatomical structures in 3-D medical images. *IEEE Transactions on Medical Imaging*, 36(7), 1470–1481. .10.1109/TMI.2017.2673121
- Deng, S.-P., Cao, S., Huang, D. S., & Wang, Y. P. (2017). Identifying stages of kidney renal cell carcinoma by combining gene expression and DNA methylation data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(5), 1147–1153. doi:10.1109/TCBB.2016.2607717 PMID:28113675
- Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S. P., Clarkson, M. J., & Barratt, D. C. (2018). Automatic multi-organ segmentation on abdominal CT with dense V-networks. *IEEE Transactions on Medical Imaging*, 37(8), 1822–1834. doi:10.1109/TMI.2018.2806309 PMID:29994628
- Han, S., Hwang, S. I., & Lee, H. J. (2019). The classification of renal cancer in 3-phase CT images using a deep learning method. *Journal of Digital Imaging*, 32(4), 638–643. doi:10.1007/s10278-019-00230-2 PMID:31098732
- Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., & Brindle, P. (2008). Predicting cardiovascular risk in England and Wales: prospective derivation and validation of qrisk2. *BMJ (Clinical Research Ed.)*, 336(7659), 1475–1482. doi:10.1136/bmj.39609.449676.25 PMID:18573856
- Jiang, J., Trundle, P., & Ren, J. (2010, December). Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 34(8), 617–631. doi:10.1016/j.compmedimag.2010.07.003 PMID:20713305
- Kakitapalli, Y., Ampolu, J., Madasu, S. D., & Sai Kumar, M. L. S. (2020). Detailed review of chronic kidney disease. *Kidney Diseases*, 6(2), 85–91. doi:10.1159/000504622 PMID:32309290
- Khalifa, N. E. M., Taha, M. H. N., Ezzat Ali, D., Slowik, A., & Hassanien, A. E. (2020). Artificial intelligence technique for gene expression by tumor RNA-seq data: A novel optimized deep learning approach. *IEEE Access: Practical Innovations, Open Solutions*, 8, 22874–22883. doi:10.1109/ACCESS.2020.2970210
- Kline, T. L., Korfiatis, P., Edwards, M. E., Blais, J. D., Czerwicz, F. S., Harris, P. C., King, B. F., Torres, V. E., & Erickson, B. J. (2017). Performance of an artificial multi-observer Deep Neural Network for Fully Automated Segmentation of Polycystic Kidneys. *Journal of Digital Imaging*, 30(4), 442–448. doi:10.1007/s10278-017-9978-1 PMID:28550374
- Kriplani, H., Patel, B., & Roy, S. (2019). *Prediction of chronic kidney diseases using deep artificial neural network technique*. .10.1007/978-3-030-04061-1\_18

- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. doi:10.1016/j.media.2017.07.005 PMID:28778026
- Lu, F., Wu, F., Hu, P., Peng, Z., & Kong, D. (2017). Automatic 3D liver location and segmentation via convolutional neural network and graph cut. *International Journal of Computer Assisted Radiology and Surgery*, 12(2), 171–182. doi:10.1007/s11548-016-1467-3 PMID:27604760
- Lu, X., Xu, D., & Liu, D. (2016). Robust 3D organ localization with dual learning architectures and fusion. *Lecture Notes in Computer Science*, 10008, 12–20. doi:10.1007/978-3-319-46976-8\_2
- Luyckx, V. A., Tonelli, M., & Stanifer, J. W. (2018). The global burden of kidney disease and the sustainable development goals. *Bulletin of the World Health Organization*, 96(6), 414–422D. 10.2471/BLT.17.206441
- Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges [Review]. *Briefings in Bioinformatics*, 19(6), 1236–1246. doi:10.1093/bib/bbx044 PMID:28481991
- Murphy Kevin, P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.
- Noll, M., Li, X., & Wesarg, S. (2014). Automated kidney detection and segmentation in 3D ultrasound. In *Proceedings of the Workshop Clin. Image-Based Procedures* (pp. 83–90). Academic Press.
- Olaf. (2016). Çiçek, Özgün and Abdulkadir, Ahmed and Lienkamp. In *U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*. Academic Press.
- Ravishankar, H., Prabhu, S. M., Vaidya, V., & Singhal, N. (2016). Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning. In *IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, 2016* (pp. 779–782). doi:10.1109/ISBI.2016.7493382
- Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvengadam, S., Annangi, P., Babu, N., & Vaidya, V. (2016). Understanding the mechanisms of deep transfer learning for medical images. *Lecture Notes in Computer Science*, 188–196. doi:10.1007/978-3-319-46976-8\_20
- Ravizza, S., Huschto, T., Adamov, A., Böhm, L., Büsser, A., Flöther, F. F., Hinzmann, R., König, H., McAhren, S. M., Robertson, D. H., Schleyer, T., Schneidinger, B., & Petrich, W. (2019). Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nature Medicine*, 25(1), 57–59. doi:10.1038/s41591-018-0239-8 PMID:30617317
- Ren, Y., Fei, H., Liang, X., Ji, D., & Cheng, M. (2019). A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. *BMC Medical Informatics and Decision Making*, 19(Suppl 2), 51. 10.1186/s12911-019-0765-4
- Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285–1298. doi:10.1109/TMI.2016.2528162 PMID:26886976
- Shon, H. S., Batbaatar, E., Kim, K. O., Cha, E. J., & Kim, K. (2020). Classification of Kidney Cancer Data Using Cost-Sensitive Hybrid Deep Learning Approach. *Symmetry*, 12(1), 154. doi:10.3390/sym12010154
- Simmons, M. N., Ching, C. B., Samplaski, M. K., Park, C. H., & Gill, I. S. (2010). Kidney tumor location measurement using the c index method. *The Journal of Urology*, 183(5), 1708–1713. doi:10.1016/j.juro.2010.01.005 PMID:20299047
- Thong, W., Kadoury, S., Piché, N., & Pal, C. J. (2018). Convolutional networks for kidney segmentation in contrast-enhanced CT scans. *Computer Methods in Biomechanics and Biomedical Engineering. Imaging & Visualization*, 6(3), 277–282. doi:10.1080/21681163.2016.1148636
- Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., & Yushkevich, P. A. (2013). Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 611–623. doi:10.1109/TPAMI.2012.143 PMID:22732662

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*, 12(4), e0174944. doi:10.1371/journal.pone.0174944 PMID:28376093

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights Into Imaging*, 9(4), 611–629. doi:10.1007/s13244-018-0639-9 PMID:29934920

Yin, S., Peng, Q., Li, H., Zhang, Z., You, X., Fischer, K., Furth, S. L., Tasian, G. E., & Fan, Y. (2020). Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks. *Medical Image Analysis*, 60, 101602. doi:10.1016/j.media.2019.101602 PMID:31760193

Zandi-Nejad, K., Luyckx, V. A., & Brenner, B. M. (2006). Adult hypertension and kidney disease the role of fetal programming. *Hypertension*, 47(3), 502–508. doi:10.1161/01.HYP.0000198544.09909.1a PMID:16415374

Zeng, Z. Q. (2008). Fast training Support Vector Machines using parallel sequential minimal optimization. In *Intelligent System and Knowledge Engineering. ISKE 2008. 3rd International Conference on Volume Abteilung I*. IEEE.


Zhang, H., Chen, Y., Song, Y., Xiong, Z., Yang, Y., & Jonathan Wu, Q. M. J. (2019). Automatic kidney lesion detection for CT images using morphological cascade convolutional neural networks. *IEEE Access*, 7, 83001–83011.

Mynapati Lakshmi Prasudha completed M.Tech from VNRVJiet and presently is doing research in the areas of Machine learning and Deep learning. Rakesh Kasumolla is an M. Tech Student at VNR VJiet. Deepak Sukheja is working as an Associate professor in the Department of Computer Science and Engineering at VNRVJiet (NAAC accredited A++ Grade, one among top 5 colleges of Hyderabad), Hyderabad, India. Deepak obtained a Ph.D. in Computer Science from Vikram University, Ujjain in 2014, Master of Technology degree in Computer Technology from Govt. Engineering College (NIT) Raipur (CG)-2005, and Master in Computer Science from (DAVV) Indore in 1999. Deepak has published more than 22 research papers, published two patent and is working as a reviewer for Elsevier Journals and others. Along with academics, Deepak has served his services as a senior software engineer to Patni Computer and Wipro Technology during the period of 2004 to 2007. His current research interests include Data Science, Deep learning, Data Analytics, Blockchain Technology, Query Optimization and Distributed System.



# Different Approaches to Reducing Bias in Classification of Medical Data by Ensemble Learning Methods

Adem Doganer, Kahramanmaraş Sutcu Imam University, Kahramanmaraş, Turkey

 <https://orcid.org/0000-0002-0270-9350>

## ABSTRACT

In this study, different models were created to reduce bias by ensemble learning methods. Reducing the bias error will improve the classification performance. In order to increase the classification performance, the most appropriate ensemble learning method and ideal sample size were investigated. Bias values and learning performances of different ensemble learning methods were compared. AdaBoost ensemble learning method provided the lowest bias value with  $n: 250$  sample size while Stacking ensemble learning method provided the lowest bias value with  $n: 500, n: 750, n: 1000, n: 2000, n: 4000, n: 6000, n: 8000, n: 10000$ , and  $n: 20000$  sample sizes. When the learning performances were compared, AdaBoost ensemble learning method and RBF classifier achieved the best performance with  $n: 250$  sample size ( $ACC = 0.956, AUC: 0.987$ ). The AdaBoost ensemble learning method and REPTree classifier achieved the best performance with  $n: 20000$  sample size ( $ACC = 0.990, AUC = 0.999$ ). In conclusion, for reduction of bias, methods based on stacking displayed a higher performance compared to other methods.

## KEYWORDS

Boosting, Deep Learning, Ensemble Learning, Machine Learning

## INTRODUCTION

Machine learning methods have been widely used in the field of data mining in recent years. Machine learning algorithms based on the theoretical structure of statistics and computer science can provide high performance in data extraction, estimation and classification. With the development of technology in the field of health, there has been a rapid increase in data. Traditional statistical methods have been insufficient in terms of performance regarding data extraction and classification. Machine learning methods have been a powerful alternative to traditional methods because they both save time and provide high performance. Machine learning methods form the basis of many artificial intelligence applications. These methods are used in many medical fields such as diagnosis, early diagnosis and pattern recognition.

DOI: 10.4018/IJBDAH.20210701.aa2

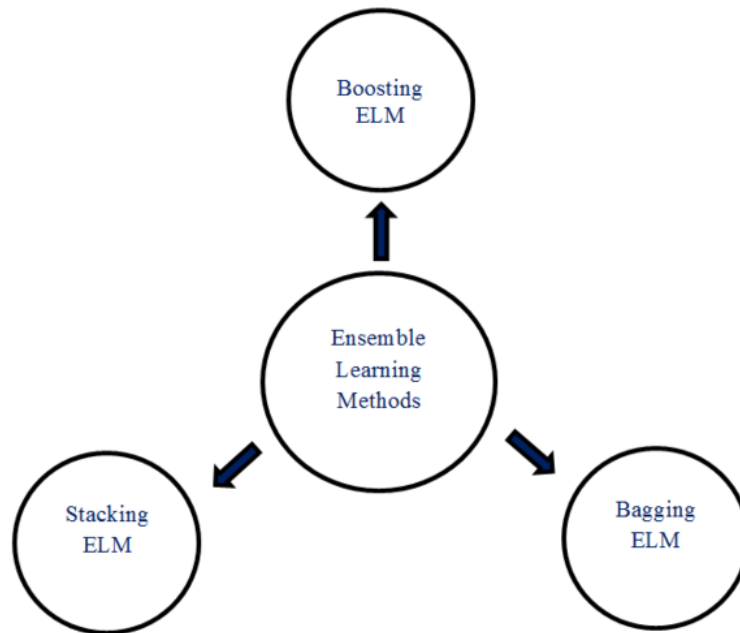
This article, published as an Open Access article on April 23, 2021 in the gold Open Access journal, International Journal of Big Data and Analytics in Healthcare (converted to gold Open Access January 1, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Although machine learning methods are widely used and can be easily learned from data, there are cases where they cannot provide high classification performance in all conditions. In some cases, although there is a high performance learning from the training data set, a poor performance is given regarding test data sets. There are different reasons for this issue. Although the model provides high accuracy performance with the training data set, the main reason for the low accuracy performance with the test data set is the problem of overfitting. The problem of overfitting is an error caused by the model's memorizing the data instead of learning the pattern in the training data set. The model that memorizes the data during the training phase provides a high accuracy performance but gives a low accuracy performance with different data due to failure to learn the pattern in the testing phase. This error, which causes the problem of overfitting, is described as variance in machine learning. Variance is not the only error in machine learning. In the training phase, the model does not provide high accuracy performance with the training data set. A model that shows low accuracy performance during the training phase will also demonstrate low accuracy performance during the testing phase. The failure of the model to achieve the desired accuracy performance in the training data set is described as the problem of underfitting. The problem of underfitting occurs when the bias error is high. Bias is an error caused by the inability of the model to learn the pattern in the training data set. There are different reasons for the occurrence of bias. One of the reasons is that the correct model has not been selected for the training data set. Some models are not sufficient to classify the data set and learn the pattern. These models are weak classifiers. Therefore, strong classifiers are used to reduce bias. However, while strong classifiers provide high performance with training data sets, they might show a lower performance with test data sets. This issue induces the problem of overfitting. Another method to reduce bias is to increase model complexity. Increasing model complexity is important for reducing bias in the training data set. However, since increasing model complexity also increases variance, it results in a poor performance with the test data set.

Numerous studies have been conducted on the reduction of bias in the classification. Brain and Gwebb (1999) investigated the effect of sample size on bias and variance in classification. They stated that increasing sample size had no effect on bias. Brain and Gwebb (2002) have examined the suitability of algorithms used for the classification of small data sets for use for bigger data sets and their effects on bias and variance. Liu et al. (2016) proposed a different approach that includes selection in order to reduce bias in machine learning and classification. Cawley and Talbot (2010) conducted a performance assessment by examining the effects on bias and variance in case of excessive learning in model selection. In their study, Hainmueller and Hazzlet (2014) proposed the Kernel-based least squares method to reduce bias in machine learning methods. By using the Naive Bayes method in their study, Yang and Webb (2009) stated that classification error could be reduced by managing bias and variance. Suen et al. (2005) compared the performances of bias and variance reduction techniques by combining them in regression trees. Lee et al. (2010) produced results with lower bias by using weight tendency scores in machine learning methods in their study. In his study, Aminian (2005) aimed to reduce bias and variance by using the Jensen-Shannon divergence. Noh et al. (2014) stated in their study that by changing the distance metric, bias can be reduced in the nearest neighbor method. Varma and Simon (2006) conducted studies on the use of cross-validation in the prediction of bias in model selection. In their study, Perdue et al. (2018) aimed to reduce model bias in deep learning classifiers by utilizing reverse neural networks.

Different methods have been developed to improve classification performance and reduce errors in machine learning. One of those methods is ensemble learning. The ensemble learning method is based on the assumption that joint estimation of estimates obtained by multiple classifiers may have a higher accuracy than the estimation of a single classifier (Zhang and Ma, 2012). In this sense, ensemble learning methods have proven to be popular and powerful method among machine learning methods. Ensemble learning methods are showed Figure 1. Wang et al. (2014) used ensemble learning methods in their study for the classification of emotions. Shi et al. (2011) used ensemble learning methods in their study for text classification. Han and Liu (2011) made use of different ensemble learning

Figure 1. Ensemble Learning Methods



methods for remote image classification. Hsieh et al. (2012) utilized ensemble learning methods in their study for early detection of breast cancer. Sun (2007) utilized ensemble learning methods in his study for the classification of EEG signals.

Ensemble learning methods are widely used for early diagnosis of diseases. Ensemble learning methods provide high classification performance. Kazemi and Mirrashondel (2018) used ensemble learning methods to predict kidney stones. Tadepalli and Lakshmi (2019) proposed a machine learning-based model for IVF prediction. Farahani et al. (2018) proposed a ensemble learning-based hybrid model to detect lung nodules from CT images. Fitriyani et al. (2019) proposed an ensemble learning-based prediction model for predicting diabetes and hypertension diseases. Wang et al. (2019) proposed a model based on stacking ensemble learning method for detection of prostate cancer.

In this study, different approaches were taken into consideration to reduce the bias observed in the machine learning methods during the training phase. There are many studies on overfitting and variance errors in the literature. However, there are very few studies aimed at reducing bias during the training phase. In our study, different methods that provide the highest performance for bias reduction were compared. In order to examine the effect of sample size on bias, data sets with different sample sizes were studied. In this study, the aim was to develop the most appropriate model in order to reduce bias and minimize the errors in the training phase. It is aimed to minimize the bias error and increase the classification performance. In order to increase the classification performance, the most appropriate ensemble learning method and ideal sample size were investigated. In this study, it is aimed to determine the most successful ensemble learning method in order to reduce the bias error. The effects of sample size on classification performance were investigated. In the model, the performances of boosting, voting stacking, bagging methods and SVM, SGD, NB, RBF, REPTREE classifiers were compared.

MATERIAL AND METHODS

Data Set

The data set of this study was produced hypothetically with simulation by considering the statistical parameters in the studies conducted on hemogram values of patients with coagulase infection and healthy individuals. Data production with simulation was carried out in accordance with the normal distribution, taking into account the mean and standard deviation parameters of the variables. Compliance of simulated data to normal distribution was checked with Q-Q graphs. 10 data sets were created in different sample sizes (n: 250, n: 500, n: 750, n: 1000, n: 2000, n: 4000, n: 6000 n: 8000, n: 10000 and n: 20000). The study included 1 target (Group) and 14 predictors (Gender, Age, Eosinophil, Monocyte, Hemoglobin, Hematocrit, Platelet, Neutrophil, Lymphocyte, RBC, WBC, CRP, MCV and IG). The definition of variables is summarized in Table 1. WEKA (Waikato Environment for Knowledge Analysis) version 3.9 was used for the evaluation of the data.

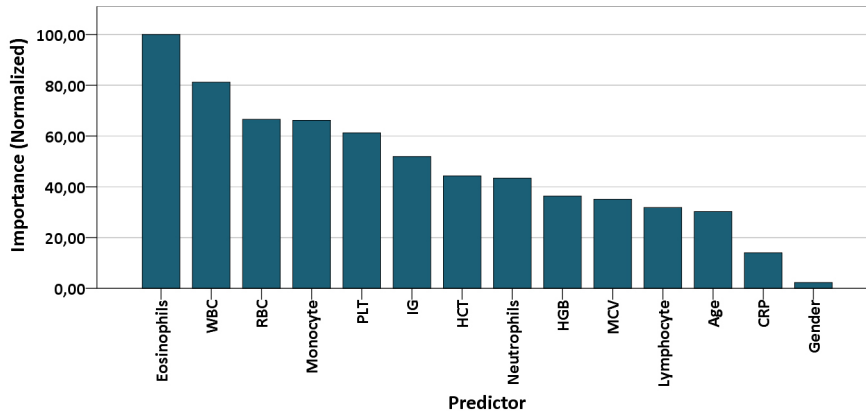
Pre-Analysis Dataset and Feature Selection

Outlier and extreme values in the data set were examined by Local Outlier Factor (LOF) algorithm. The LOF Algorithm scales the distance of objects close to it for each data depending on the local densities of the data. It is a powerful method used to detect local outlier, outlier and outlier observations (Breunig et al.2000; Lee et al. 2011). Outlier and extreme values were excluded from the study. Standardization was applied to quantitative data. In order to provide the best classification performance and to reduce the noisy data by identifying the variables that are not related to the model, feature selection was applied to the predictor variables. As a result of feature selection, CRP and gender variable, which had the least contribution in explaining the target variable, were excluded from the model. The feature selection process is shown in Figure 2.

Table 1. The definition of the Variables in the Current Study

Variables	Definition	Role
Group	Patient/Control	Target
Gender	Male/Female	Predictor
Eosinophil	Integer	Predictor
Monocytes	Integer	Predictor
Hemoglobin	Integer	Predictor
Hematocrit	Integer	Predictor
Platelets	Integer	Predictor
Neutrophil	Integer	Predictor
Lymphocytes	Integer	Predictor
RBC	Integer	Predictor
WBC	Integer	Predictor
Age	Integer	Predictor
CRP	Integer	Predictor
MCV	Integer	Predictor
IG	Integer	Predictor

Figure 2. Importance Values of Predictor Variables



## Machine Learning Experimentation Methods

In this study, 10 data sets consisting of different sample sizes were trained to evaluate the classification performances with different classifiers and ensemble methods. Data sets were trained with Support vector Machine (SVM), Stochastic Gradient Descent (SGD), Radial Basis Function Classifier (RBF), Naive Bayes, REPTree, Random Forest and K-NN classifiers. Random Forest and K-NN classifiers were excluded from the model because they had overfitting problems. The performances and bias rates of Bagging, Boosting, Voting and Stacking ensemble methods were evaluated in addition to the performances of each basic classifier in the model. AdaBoost method was applied for Boosting ensemble method. Random Forest classifier was utilized as meta classifier in the Stacking method. In the comparison of the bias rates, the mean of the bias rates of the classifiers in the non-ensemble method and the bias rates of the Boosting, Bagging ensemble methods were compared. Data sets were trained with K-10 fold cross-validation and by division into 70% training and 30% testing. Performance comparisons were performed with Accuracy, Precision, Recall, AUC and Kappa metrics. The bias values of the ensemble learning methods during training were also compared. The confusion matrix for metrics are given Table 2. Main model of study and process are shown in Figure 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

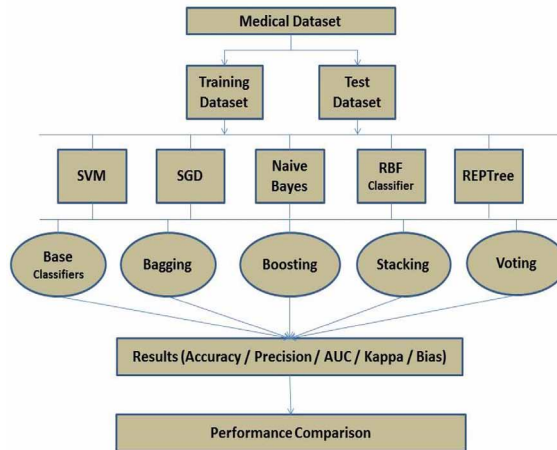
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Table 2. Confusion Matrix

		Actual Condition	
		Positive	Negative
Test Result	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 3. Classifiers, Ensemble Methods and Main Model of Study



## Bagging

Bagging (Bootstrap Aggregation), one of the first ensemble learning methods, is a simple and effective method. Developed by Breiman (Breiman, 1996) this method is based on the training of different subsets of a data set resampled by Bootstrap method in parallel with separate classifiers and testing in the same data set. The ensemble decision is determined by applying the majority voting method to the estimates obtained from each classifier. (Zhang and Ma, 2012).

## Boosting

Boosting is a powerful ensemble method based on weighting that offers a more precise and powerful classification chance by focusing on iterative steps and errors in the previous training at each step (Schapire, 1990). Although it is structurally similar to the Bagging method, in Bagging method, the majority decision is determined by the majority vote for the estimates produced in separate classifiers at the same time. In the Boosting method, the processes proceed in an iterative way and not simultaneously. In each iterative process, a strong classifier is obtained by Boosting method in order to prevent the same errors from occurring by considering the errors in the previous estimation. (Zhou, 2012). One of the most powerful Boosting algorithms is the AdaBoost algorithm. AdaBoost algorithm minimizes the exponential loss function (Freund and Schapire, 1996).

## Voting

Voting, one of the widely used ensemble learning methods, is based on combining estimates obtained from different classifiers. In the Voting process, several different classifiers train the same data set in parallel. The Voting process is applied to all estimates obtained by each classifier and the estimate of the majority is accepted as the estimate of the ensemble (Hansen and Salamon, 1990; Zhang and Ma, 2012).

## Stacking

The Stacking ensemble learning method accepts estimates that different classifiers obtain from the training data set as input for a meta classifier. These input data, consisting of estimates, are re-trained in the meta classifier to obtain ensemble estimation (Wolpert, 1992; Zhou, 2012).

## Classifiers

### *Support Vector Machine*

It is a powerful classifier that is widely used in classification processes. In the linear plane, two hyperlines are formed as a boundary to separate the data belonging to the two groups. These lines aim to keep the distance between the lines at the highest level in an optimal way to separate the data of the two groups. It is a supervised machine learning algorithm that produces successful results for linear and nonlinear classification problems. Hyperparameter optimization is performed to ensure the best classification values. Support vector machines can be implemented with different core functions (Vapnik, 2013; Cortes and Vapnik, 1995).

### *Stochastic Gradient Descent*

Stochastic gradient descent applies a classification process that supports loss functions and penalties in linear models. Its successful performance with high data sizes makes this classifier popular. While a normal gradient descent model considers all values in the data set and iteratively works on the entire data set when changing weight, in the stochastic gradient descent model a single point is taken into account. This provides rapidity compared to other methods (Bottou, 2010; Çürük et al. 2018).

### *Naive Bayes*

Naive Bayes is a supervised classifier based on probabilistic calculations grounded on Bayes theorem. It can produce successful results in big data. The properties in the data set are independent of each other. The Naive Bayes classifier estimates which class the samples in the data set are included in (Rish, 2001). Naive Bayes classifier is very successful in the classification of categorical data such as text mining (Serrano-Guerrero et al. 2015).

### *Radial Basis Function Classifier*

The radial basis function classifier is based on radial-based function neural networks and provides the least squares error in the optimization process using the BFGS method. In this method that performs supervised learning, normalization is applied to all features [0,1] (Frank, 2014).

### *Reduced Error Pruning Tree (REPTree)*

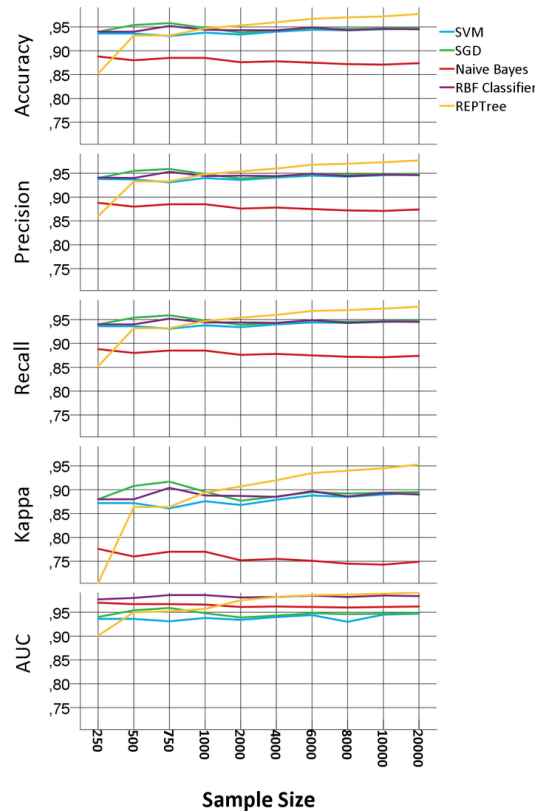
The Reduced Error Pruning Tree (REPTree) classifier determines the best tree among the many decision trees it produces with different iterations. The average squares error generated by the tree is used as a measure for pruning with the aim of estimation. Knowledge acquisition and variance are used to classify and form a decision tree. One of the most important aspects of this classifier is that it is a fast classifier. (Srinivasan and Mekala, 2014; Kalmegh, 2015).

## Results

In the first stage, the performances of 10 data sets with different sample sizes were evaluated in base classifiers without using any ensemble learning methods. Accuracy, Precision, Recall, AUC and Kappa metrics were compared. The findings of the comparison are given in Figure 4. According to the findings in Figure 4, SGD classifier provided the highest performance with n: 250 sample size in accuracy metric. As for the n: 20000 sample size, the REPTree classifier provided the highest performance. In all metrics, the REPTree algorithm showed a performance increase parallel to the increase in sample size. The performances of the classifiers according to the sample sizes are shown in Figure 4.

In addition to the performances of base classifiers, their performances with ensemble methods were also evaluated in the study. accuracy performances of classifiers in different sample sizes with the Boosting (AdaBoost), Bagging ensemble methods and without any ensemble method were compared. The highest performance in terms of accuracy metric was observed in the AdaBoost (Boosting)

Figure 4. Performances of Base Classifiers at Different Sample Sizes



ensemble method. AdaBoost method significantly improved performance for Naive Bayes, RBF and REPTree classifiers. The ensemble methods did not provide a significant increase for SVM and SGD classifiers. The highest performance was observed in the REPTree classifier in the AdaBoost ensemble learning method. Additionally, increasing sample size in REPTree classifier contributed to increase in performance. The accuracy performances of ensemble methods and classifiers at different sample sizes are shown in Figure 5. In terms of Precision, Recall and Kappa metrics, the performance results of the methods and classifiers are similar to the accuracy metric results. Again, the best performance in these metrics was achieved with the AdaBoost ensemble learning method. The performance of methods and classifiers in terms of Precision, Recall and Kappa metrics are shown in Figure 6, Figure 7 and Figure 8 respectively. In terms of AUC metrics, the AdaBoost ensemble learning method was the best performing method for all classifiers. The highest performance was achieved with n: 20000 sample size with the REPTree classifier and the AdaBoost ensemble method. The performances of the methods and classifiers in terms of AUC metric are given in Figure 9.

In this study, classification performances of ensemble learning methods at n: 250 and n: 20000 sample sizes were evaluated. The highest classification performance in the n: 250 sample size was demonstrated by the RBF classifier with AdaBoost ensemble learning method. The highest classification performance in the n: 20000 sample size was demonstrated by the REPTree classifier with AdaBoost ensemble learning method. The classification performances of the methods are shown in Table 3.

The main findings of the study are the bias values that emerged in the training data set of ensemble learning methods. At this stage, the bias values the ensemble learning methods revealed



Figure 5. Accuracy Performances of Ensemble Methods

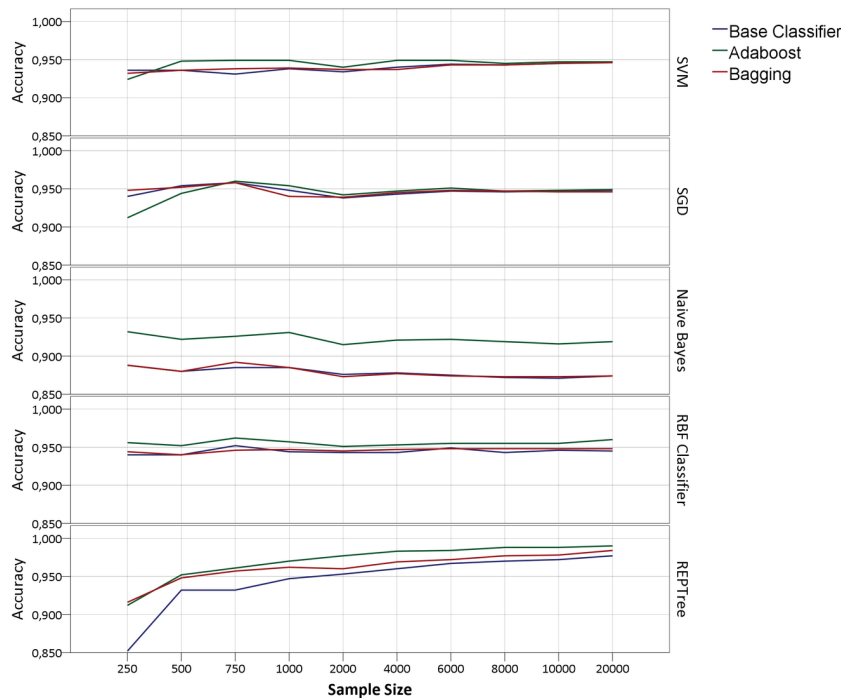


Figure 6. Precision Performances of Ensemble Methods

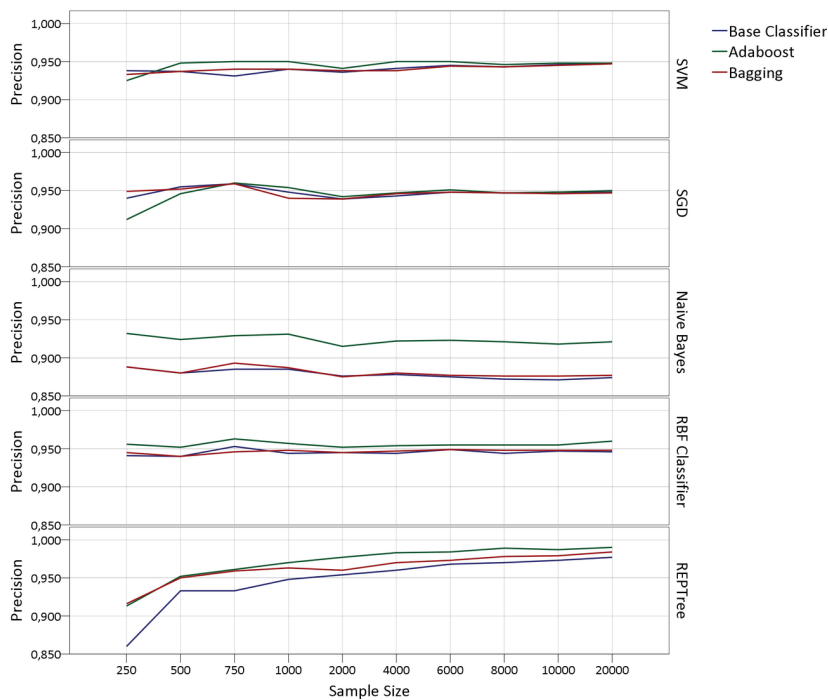


Figure 7. Recall Performances of Ensemble Methods

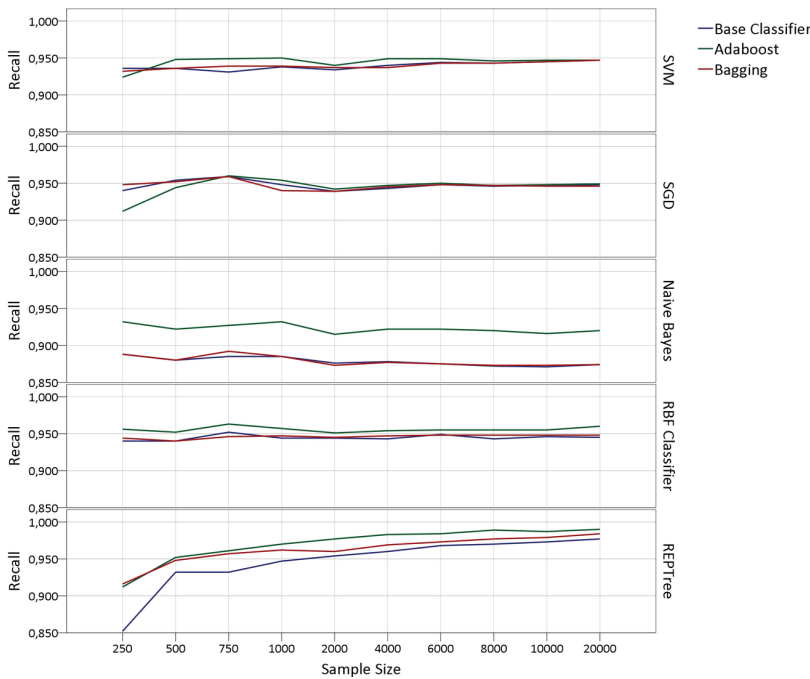


Figure 8. Kappa Performances of Ensemble Methods

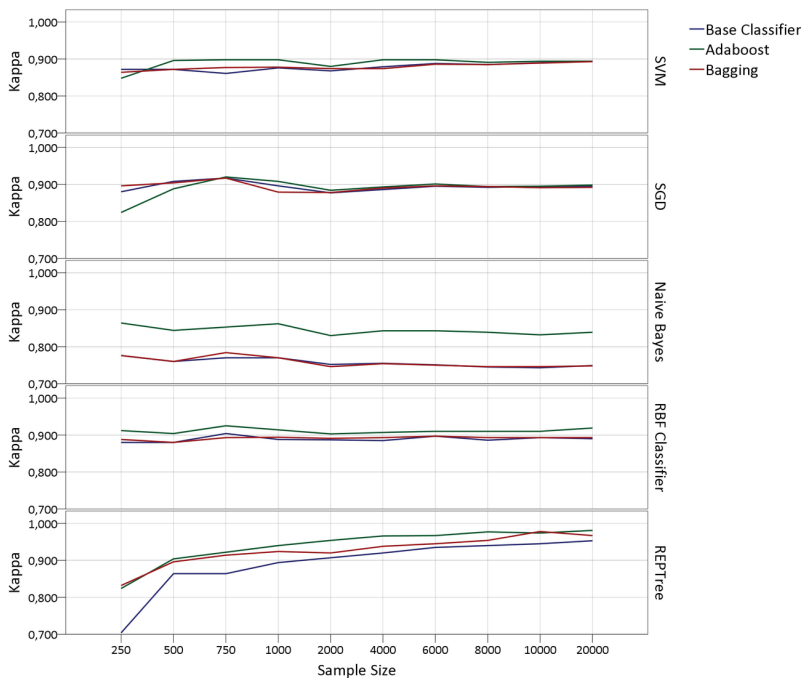


Figure 9. AUC Performances of Ensemble Methods

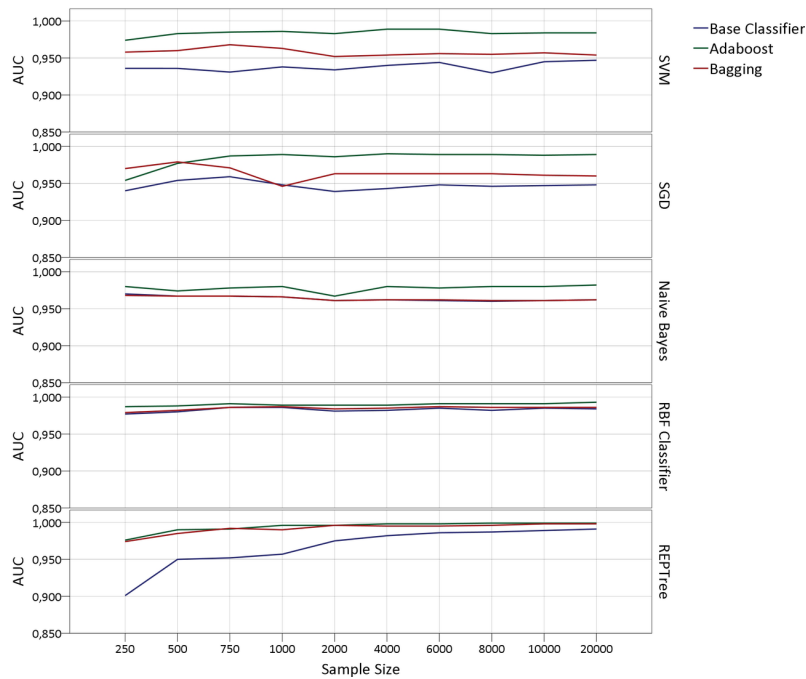


Table 3. All metrics performances of ensemble methods at n:250 and n:20000 sample sizes

		n: 250					n:20000				
		Acc.	Prec.	Recall	AUC	Kappa	Acc.	Prec.	Recall	AUC	Kappa
Base Classifier	SVM	0,936	0,938	0,936	0,936	0,872	0,947	0,948	0,947	0,947	0,894
	SGD	0,940	0,940	0,940	0,940	0,880	0,948	0,948	0,948	0,948	0,895
	Naive Bayes	0,888	0,888	0,888	0,970	0,776	0,874	0,874	0,874	0,962	0,749
	RBF Classifier	0,940	0,941	0,940	0,977	0,880	0,945	0,946	0,945	0,984	0,890
	REPTree	0,852	0,860	0,852	0,901	0,704	0,977	0,977	0,977	0,991	0,953
Adaboost	SVM	0,924	0,925	0,924	0,974	0,848	0,947	0,948	0,947	0,984	0,894
	SGD	0,912	0,912	0,912	0,954	0,824	0,949	0,950	0,949	0,989	0,898
	Naive Bayes	0,932	0,932	0,932	0,980	0,864	0,919	0,921	0,920	0,982	0,839
	RBF Classifier	<b>0,956</b>	<b>0,956</b>	<b>0,956</b>	<b>0,987</b>	<b>0,912</b>	0,960	0,960	0,960	0,993	0,919
	REPTree	0,912	0,913	0,912	0,976	0,824	<b>0,990</b>	<b>0,990</b>	<b>0,990</b>	<b>0,999</b>	<b>0,981</b>
Bagging	SVM	0,932	0,933	0,932	0,958	0,864	0,946	0,947	0,947	0,954	0,893
	SGD	0,948	0,949	0,948	0,970	0,896	0,946	0,947	0,946	0,960	0,892
	Naive Bayes	0,888	0,888	0,888	0,968	0,776	0,874	0,877	0,874	0,962	0,748
	RBF Classifier	0,944	0,945	0,944	0,979	0,888	0,948	0,948	0,948	0,986	0,893
	REPTree	0,916	0,916	0,916	0,974	0,832	0,984	0,984	0,984	0,998	0,967
Stacking		0,944	0,944	0,944	0,975	0,888	0,978	0,978	0,978	0,994	0,955
Voting		0,944	0,945	0,945	0,978	0,888	0,955	0,955	0,955	0,992	0,909

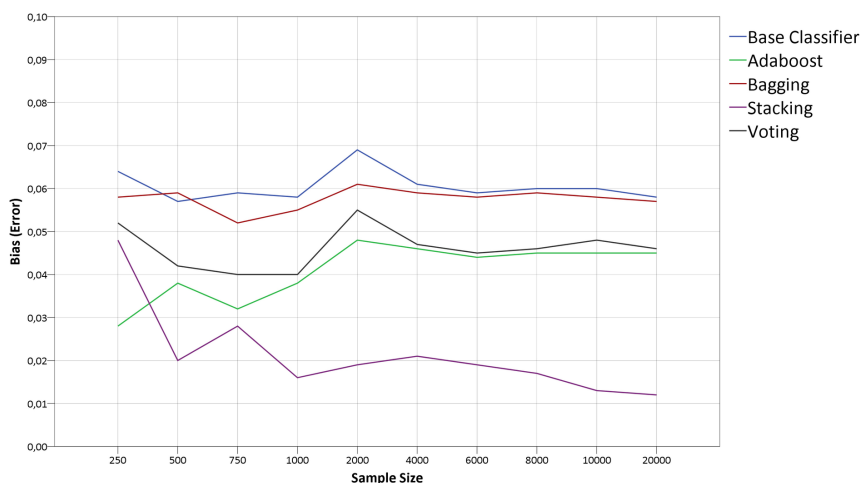
in the training data set were compared. In the comparison, the mean values of bias obtained from single classifiers and classifiers in the AdaBoost and bagging methods were calculated. According to the comparison results, the lowest bias (bias error) value was obtained by AdaBoost method with  $n$ : 250 sample size. On the other hand, in all other sample sizes ( $n$ : 500,  $n$ : 750,  $n$ : 1000,  $n$ : 2000,  $n$ : 4000,  $n$ : 6000,  $n$ : 8000,  $n$ : 10000,  $n$ : 20000), the lowest bias value was provided by the Stacking ensemble learning method. It was observed that bias value decreased with increasing sample size in the Stacking ensemble learning method. The performances of the ensemble methods with their bias values are given in Figure 10.

## DISCUSSION

The development of technology has enabled significant achievements to be achieved in human health. Technology has made people's lives easier. Diseases could be early diagnosed (So et al. 2017; Parisi et al. 2018). Access to health information has become easier (Beam and Kohane, 2018). New treatment methods have been developed (Scheeder et al 2018). Telemedicine services and home nursing services have reached even more widespread use (Ramkumar et al. 2018). Important improvements have been made in elderly care and elderly health (Özsungur, 2019a; Özsungur, 2019b). The accuracy performance of medical imaging reports and laboratory results has increased (Suzuki, 2017). Artificial intelligence and machine learning methods form the basis of technological developments in the health system.

Behind the artificial intelligence technologies that have been rapidly in recent years are powerful machine learning methods. The high performance of classifiers plays an important role in the popularity of machine learning methods. The first condition for the high success of classification in machine learning methods depends on minimizing the bias in the training data sets of classification algorithms and the model learning the pattern correctly. Different methods have been developed in order to have high classification performance in machine learning methods. Some of those methods are ensemble learning methods. Ensemble learning methods are based on the principle that the performance of multiple classifiers will be better than the performance of a single classifier. In their study, Wang et al. (2011) compared the performances of the credit scores of different countries by classifying them with basic classifiers, Bagging, Boosting and Stacking ensemble methods. Accuracy, type 1 and type 2 error metrics were evaluated in the comparison. They achieved the best performance with Bagging

**Figure 10. Bias Performances of Ensemble Methods**



and Stacking ensemble methods and Decision Tree classifier. They stated that the Bagging method showed a better performance than the Boosting method. In our study, the best performances were provided by RBF classifier in the Boosting ensemble method and REPTree classifier in large sample size. Tan and Gilbert (2003) compared the performance of C4.5 basic classifier, Bagging and Boosting ensemble learning methods for cancer classification in their study. In the study, it was observed that Bagging and Boosting ensemble methods outperformed the basic classifiers. In our study, ensemble methods provided higher accuracy performance as well. Sun (2002) used ensemble learning methods to predict sound accents in a study. The ensemble learning methods were observed to provide better results than basic learners. In their study, Kaur and Kaur (2014) compared the performances of different classifiers in ensemble learning methods. As a result of the comparison, they observed that ensemble methods increased the performance compared to the basic classifiers. Das and Sengur (2010) compared the performances of ensemble learning methods in the diagnosis of heart disorders in their study. They stated that AdaBoost ensemble method provides higher accuracy performance than basic classifier, Bagging and random Subspace ensemble methods. The AdaBoost method provided the highest accuracy according to the findings of our study as well. Our study demonstrated the effect of ensemble learning methods on bias in training data sets. In their study, Liu and Cocea (2017) made a recommendation based on granular calculation to reduce bias in ensemble methods. Lu et al. (2017) used the ensemble learning methods in their study to reduce the bias in the new computer-enhanced diagnosis systems used for the detection of breast cancer.

## CONCLUSION

In our study, the ensemble learning methods were observed to reduce bias in the training data set. Stacking ensemble learning method has been identified as the most successful method in reducing bias. On the other hand, increasing sample size in Voting, Bagging and Boosting ensemble methods does not contribute to decreasing bias. It has been observed that increasing sample size in the Stacking ensemble method contributes to reducing bias.

Ensemble learning methods make a significant contribution to reducing bias in accuracy performance and training data sets when used with appropriate data and appropriate classifiers. The use of the Stacking ensemble learning method contributes significantly to the reduction of bias. Not each classifier or ensemble method provides an increase in performance despite increasing sample size. The AdaBoost (boosting) ensemble learning method provides high performance with RBF classifier in small sample size data set and REPTree classifier in large sample size data set in accuracy performance.

In future studies, it is planned to create hybrid models based on the stacking ensemble learning method. Hybrid models based on ensemble learning model can be applied for early diagnosis of diseases. Ensemble learning-based hybrid models are predicted to be classified with high performance.

## ACKNOWLEDGMENT

This study was presented as an oral Presentation in 11 International Statistics Congress (4-8 October 2019). Data on humans and animals were not used in this study. The data were hypothetically generated by simulation. Therefore, ethics committee approval was not required. No funding for this study.

## REFERENCES

- Aminian, A. (2005). Active learning for reducing bias and variance of a classifier using Jensen-Shannon divergence. In *Fourth International Conference on Machine Learning and Applications (ICMLA'05)* (pp. 6-pp). IEEE. doi:10.1109/ICMLA.2005.7
- Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Journal of the American Medical Association*, 319(13), 1317–1318. doi:10.1001/jama.2017.18391 PMID:29532063
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT, 2010*, 177–186. doi:10.1007/978-3-7908-2604-3\_16
- Brain, D., & Webb, G. (1999). On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales* (pp. 117-128). Academic Press.
- Brain, D., & Webb, G. I. (2002, August). The need for low bias algorithms in classification learning from large data sets. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 62-73). Springer. doi:10.1007/3-540-45681-3\_6
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. doi:10.1007/BF00058655
- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104). doi:10.1145/342009.335388
- Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul), 2079–2107.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. doi:10.1007/BF00994018
- Çürük, E., Acı, Ç., & Eşsiz, E. S. (2018, September). Performance Analysis of Artificial Neural Network Based Classifiers for Cyberbullying Detection. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)* (pp. 1-5). IEEE. doi:10.1109/UBMK.2018.8566566
- Das, R., & Sengur, A. (2010). Evaluation of ensemble methods for diagnosing of valvular heart disease. *Expert Systems with Applications*, 37(7), 5110–5115. doi:10.1016/j.eswa.2009.12.085
- Farahani, F. V., Ahmadi, A., & Zarandi, M. H. F. (2018). Hybrid intelligent approach for diagnosis of the lung nodule from CT images using spatial kernelized fuzzy c-means and ensemble learning. *Mathematics and Computers in Simulation*, 149, 48–68. doi:10.1016/j.matcom.2018.02.001
- Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *IEEE Access: Practical Innovations, Open Solutions*, 7, 144777–144789. doi:10.1109/ACCESS.2019.2945129
- Frank, E. (2014). *Fully supervised training of Gaussian radial basis function networks in WEKA* (Computer Science Working Papers, 04/2014). Hamilton, New Zealand: Department of Computer Science, The University of Waikato.
- Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *ICML (Vol. 96, pp. 148-156)*. Academic Press.
- Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2), 143–168. doi:10.1093/pan/mpt019
- Han, M., & Liu, B. (2015). Ensemble of extreme learning machine for remote sensing image classification. *Neurocomputing*, 149, 65–70. doi:10.1016/j.neucom.2013.09.070
- Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10), 993–1001.

- Hsieh, S. L., Hsieh, S. H., Cheng, P. H., Chen, C. H., Hsu, K. P., Lee, I. S., Wang, Z., & Lai, F. (2012). Design ensemble machine learning model for breast cancer diagnosis. *Journal of Medical Systems*, 36(5), 2841–2847. doi:10.1007/s10916-011-9762-6 PMID:21811801
- Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. *International Journal of Innovative Science. Engineering & Technology*, 2(2), 438–446.
- Kaur, A., & Kaur, K. (2014, September). Performance analysis of ensemble learning for predicting defects in open source software. In 2014 international conference on advances in computing, communications and informatics (ICACCI) (pp. 219–225). IEEE. doi:10.1109/ICACCI.2014.6968438
- Kazemi, Y., & Mirroshandel, S. A. (2018). A novel method for predicting kidney stone type using ensemble learning. *Artificial Intelligence in Medicine*, 84, 117–126. doi:10.1016/j.artmed.2017.12.001 PMID:29241659
- Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3), 337–346. doi:10.1002/sim.3782 PMID:19960510
- Lee, J., Kang, B., & Kang, S. H. (2011). Integrating independent component analysis and local outlier factor for plant-wide process monitoring. *Journal of Process Control*, 21(7), 1011–1021. doi:10.1016/j.jprocont.2011.06.004
- Liu, H., & Cocea, M. (2017). Granular computing-based approach for classification towards reduction of bias in ensemble learning. *Granular Computing*, 2(3), 131–139.
- Liu, H., Gegov, A., & Cocea, M. (2016). Nature and biology inspired approach of classification towards reduction of bias in machine learning. In 2016 International Conference on Machine Learning and Cybernetics (ICMLC) (Vol. 2, pp. 588–593). IEEE. doi:10.1109/ICMLC.2016.7872953
- Lu, W., Li, Z., & Chu, J. (2017). A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning. *Computers in Biology and Medicine*, 83, 157–165. doi:10.1016/j.combiomed.2017.03.002 PMID:28282591
- Noh, Y. K., Sugiyama, M., Liu, S., Plessis, M. C., Park, F. C., & Lee, D. D. (2014, April). Bias reduction and metric learning for nearest-neighbor estimation of Kullback-Leibler divergence. In Artificial Intelligence and Statistics (pp. 669–677). Academic Press.
- Özşungur, F. (2019a). Gerontechnological factors affecting successful aging of elderly. *The Aging Male*, 1–13. doi:10.1080/13685538.2018.1539963 PMID:30741066
- Özşungur, F. (2019b). A Research on the Effects of Successful Aging on the Acceptance and Use of Technology of the Elderly. *Assistive Technology*, 1–14. doi:10.1080/10400435.2019.1691085 PMID:31710261
- Parisi, L., RaviChandran, N., & Manaog, M. L. (2018). Feature-driven machine learning to improve early diagnosis of Parkinson's disease. *Expert Systems with Applications*, 110, 182–190.
- Perdue, G. N., Ghosh, A., Wospakrik, M., Akbar, F., Andrade, D. A., Ascencio, M., & Cai, T. et al. (2018). Reducing model bias in a deep learning classifier using domain adversarial neural networks in the MINERvA experiment. *Journal of Instrumentation: An IOP and SISSA Journal*, 13(11), P11020. doi:10.1088/1748-0221/13/11/P11020
- Ramkumar, P. N., Haeberle, H. S., Navarro, S. M., Sultan, A. A., Mont, M. A., Ricchetti, E. T., Schickendantz, M. S., & Iannotti, J. P. (2018). Mobile technology and telemedicine for shoulder range of motion: Validation of a motion-based machine-learning software development kit. *Journal of Shoulder and Elbow Surgery*, 27(7), 1198–1204. doi:10.1016/j.jse.2018.01.013 PMID:29525490
- Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41–46). Academic Press.
- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2), 197–227. doi:10.1007/BF00116037
- Scheeder, C., Heigwer, F., & Boutros, M. (2018). Machine learning and image-based profiling in drug discovery. *Current Opinion in Systems Biology*, 10, 43–52. doi:10.1016/j.coisb.2018.05.004 PMID:30159406
- Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311, 18–38. doi:10.1016/j.ins.2015.03.040

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2 • July-December 2021

- Shi, L., Ma, X., Xi, L., Duan, Q., & Zhao, J. (2011). Rough set and ensemble learning based semi-supervised algorithm for text classification. *Expert Systems with Applications*, 38(5), 6300–6306. doi:10.1016/j.eswa.2010.11.069
- So, A., Hooshyar, D., Park, K. W., & Lim, H. S. (2017). Early diagnosis of dementia from clinical data by machine learning techniques. *Applied Sciences (Basel, Switzerland)*, 7(7), 651. doi:10.3390/app7070651
- Srinivasan, D. B., & Mekala, P. (2014). Mining social networking data for classification using reptree. *International Journal of Advance Research in Computer Science and Management Studies*, 2(10).
- Suen, Y. L., Melville, P., & Mooney, R. J. (2005). Combining bias and variance reduction techniques for regression trees. In *European Conference on Machine Learning* (pp. 741-749). Springer. doi:10.1007/11564096\_76
- Sun, S. (2007, May). Ensemble learning methods for classifying EEG signals. In *International Workshop on Multiple Classifier Systems* (pp. 113-120). Springer. doi:10.1007/978-3-540-72523-7\_12
- Sun, X. (2002). Pitch accent prediction using ensemble machine learning. In *Seventh international conference on spoken language processing*. Academic Press.
- Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological Physics and Technology*, 10(3), 257–273. doi:10.1007/s12194-017-0406-5 PMID:28689314
- Tadepalli, S. K., & Lakshmi, P. V. (2019). Application of Machine Learning and Artificial Intelligence Techniques for IVF Analysis and Prediction. *International Journal of Big Data and Analytics in Healthcare*, 4(2), 21–33. doi:10.4018/IJBDAH.2019070102
- Tan, A. C., & Gilbert, D. (2003). *Ensemble machine learning on gene expression data for cancer classification*. Academic Press.
- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1), 91. doi:10.1186/1471-2105-7-91 PMID:16504092
- Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1), 223–230. doi:10.1016/j.eswa.2010.06.048
- Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, 57, 77–93. doi:10.1016/j.dss.2013.08.002
- Wang, Y., Wang, D., Geng, N., Wang, Y., Yin, Y., & Jin, Y. (2019). Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Applied Soft Computing*, 77, 188–204. doi:10.1016/j.asoc.2019.01.015
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259. doi:10.1016/S0893-6080(05)80023-1 PMID:18276425
- Yang, Y., & Webb, G. I. (2009). Discretization for naive-Bayes learning: Managing discretization bias and variance. *Machine Learning*, 74(1), 39–74. doi:10.1007/s10994-008-5083-5
- Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble machine learning: methods and applications*. Springer Science & Business Media. doi:10.1007/978-1-4419-9326-7
- Zhou, Z. H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC. doi:10.1201/b12207

Adem Doganer has a Bachelor's degree in Statistics, Master's degree in Statistical Information Systems and PhD on Statistical Information Systems and Modeling. He has a second PhD on Biostatistics and Medical Informatics. Dr Adem Doganer is a faculty member in Faculty of Medicine at Kahramanmaraş Sutcu Imam University. He has journal articles about statistics, artificial intelligence and data mining. He is statistics editor of KSU Medical Journal.



# ICTs and Domestic Violence (DV): Exploring Intimate Partner Violence (IPV)

Bolanle A. Olaniran, Texas Tech University, USA

## ABSTRACT

The use of information communication technologies (ICTs) to empower individuals through social support, help-seeking, and help-providing activities is finding its place in healthcare delivery. ICTs, in particular, offer access to timely and relevant information that domestic violence victims and organizations can tap into. Thus, this article explores the use of ICTs for providing and facilitating support and care-giving services to victims/survivors of domestic violence with online communities and other groups.

## KEYWORDS

Caregivers, Domestic Violence, Healthcare, Information Communication Technologies, Intimate Partner Violence, Migrant Women Workers

## INTRODUCTION

As a communications medium, Computer-Mediated Communication (CMC) could connect geographically dispersed individuals without the constraints of time or space. Thus, individuals with diverse backgrounds, experiences, and ethnicities can share information and communicate with other individuals or groups (e.g., online community) at one time over the Internet. Online communities—where individuals with similar interests and/or experiences come together to interact—can benefit from CMC as a tool for seeking, gaining, and sharing knowledge and experiences. It is these communities—groups of individuals with similar interests and experiences who are connected using information communication technologies (ICTs) and whose conversations are facilitated through CMC use—that makes these ICTs a valuable tool for social support. Thus, this article proposes, the need to explore the use of ICTs; specifically, the role of CMC as a support medium for victims/survivors of domestic violence (DV). DV is an issue critical to healthcare and the overall general well-being of women, their families, and societies in general (Olaniran & Rodriguez, 2013).

## BACKGROUND ON DOMESTIC VIOLENCE

ICTs offer access to timely and relevant information, which allows DV organizations to serve as advocates and respond to specific cases of abuse (Hamm, 2001). Online DV organizations also provide other advocates with health and support information in order to better facilitate and provide a solution to victims/survivors of DV and other related types of sensitive healthcare issues (Campbell, Sy, and

DOI: 10.4018/IJBDAH.20210701.0a3

This article, published as an Open Access article on April 23, 2021 in the gold Open Access journal, International Journal of Big Data and Analytics in Healthcare (converted to gold Open Access January 1, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Anderson, 2000; NCIPC, 2003). The amount of information available online is also used to provide online training for individuals and organizations that deal with violence against women (VAW). A specific focus in this paper is the use of ICTs for providing and facilitating support and care giving services to victims/survivors of DV. Traditionally, ICT use in healthcare and healthcare delivery, primarily focus on hospital settings such as health care gives interaction with one another and patients or pharmacies and other agencies such that issues surrounding telemedicine and informatics are a commonplace. Notwithstanding, the Center for Disease and Control (CDC) classifies domestic violence as a serious *public health issue* (2008). According to CDC (2017), domestic violence is a serious, yet preventable, public health problem affecting more than 32 million Americans—over 10% of the United States population (CDC, 2008, 2017). The intimate partner violence (IPV) alone affecting more than 12 million people each year. Women are disproportionately impacted (CDC, 2017).

### National Domestic Violence Statistics

- 1 in 4 women and 1 in 7 men will experience severe physical violence by an intimate partner in their lifetime (CDC, 2017).
- 1 in 10 women in the United States will be raped by an intimate partner in her lifetime.
- An estimated 9.7% of women and 2.3% of men have been stalked by an intimate partner during their lifetime (CDC, 2017).

### Other Domestic Violence Data

- Female victims of intimate partner violence experienced different patterns of violence than male victims.
- Female victims experienced multiple forms of these types of violence, male victims most often experienced physical violence. Most of this victimization starts early in life.
- Approximately 80% of female victims experienced their first rape before the age of 25 and almost half experienced the first rape before age 18 (30% between 11-17 years old and 12% at or before the age of 10).
- About 35% of women who were raped as minors were also raped as adults compared to 14% of women without an early rape history.
- 28% of male victims of rape were first raped when they were 10 years old or younger (see CDC, 2017).

Studies have also demonstrated the impact of intimate partner violence on the endocrine and immune systems through chronic stress or other mechanisms (Breiding, Black, & Ryan, 2008; Crofford, 2007; Leserman & Drossman, 2007). The problems include but not limited to Fibromyalgia, Irritable bowel syndrome, Gynecological disorders; Pregnancy difficulties such as low birth weight babies and prenatal deaths, sexually transmitted diseases including HIV/AIDS, Central nervous system disorders, Heart or circulatory conditions. Intimate partner violence (IPV)—whether sexual, physical, or psychological—leads to various long-term chronic disease and other health issues that exhibit PTSD symptoms (CDC, 2017; Roberts, Klein, & Fisher, 2003).

### TYPES OF DOMESTIC VIOLENCE

According to Saltzman, Fanslow, McMahon, and Shelley (2002), there are four major categories of DV, including:

- *Physical violence* is the intentional use of physical force with the potential for causing death, disability, injury, or harm. Physical violence includes, but is not limited to, scratching; pushing;

shoving; throwing; grabbing; biting; choking; shaking; slapping; punching; burning; use of a weapon; and use of restraints or one's body, size, or strength against another person.

- *Sexual violence* is divided into three categories: 1) Use of physical force to compel a person to engage in a sexual act against his or her will, whether or not the act is completed; 2) Attempted or completed sex act involving a person who is unable to understand the nature or condition of the act, to decline participation, or to communicate unwillingness to engage in the sexual act, e.g., because of illness, disability, or the influence of alcohol or other drugs, or because of intimidation or pressure; and, 3) Abusive sexual contact.
- *Threats of physical or sexual violence* use words, gestures, or weapons to communicate the intent to cause death, disability, injury, or physical harm.
- *Psychological/emotional violence* involves trauma to the victim caused by acts, threats of acts, or coercive tactics. Psychological/emotional abuse can include, but is not limited to, humiliating the victim, controlling what the victim can and cannot do, withholding information from the victim, deliberately doing something to make the victim feel diminished or embarrassed, isolating the victim from friends and family, and denying the victim access to money or other basic resources. It is considered psychological/emotional violence when there has been prior physical or sexual violence or prior threats of physical or sexual violence.

In addition, stalking is often included among the types of IPV. Stalking generally refers to harassment or threats including: –following or appearing at a person's home/residence or place of business, harassing phone calls, leaving written messages or objects on an individual's property, and vandalizing properties (Hien & Ruglass, 2008).

## Social Networks and DV

DV support groups and services on social networks and new media are a unique type of community, where individuals who have experienced DV can understand, access support, and interact with others—particularly victims/survivors—about the issues surrounding DV. DV organizations, activists, and support groups have begun to—and will continue to increase services, goods, information, and the support they provide via ICTs. At the same time, the number of women who experience DV continues to increase each year (Woodluck, 2016). Consequently, this is a very serious and continuing challenge to public health. DV is a serious matter of concern that needs, and certainly deserves, further inquiry. The role of ICTs in providing/facilitating support for victims/survivors is important in understanding this potentially dangerous healthcare problem.

Research suggests that being involved with or living with an abusive partner can have a profound impact on a woman's general and mental health (El Morr & Layal, 2019; Longo, 2018; Simmons, Lindsey, Delaney, Whalley, & Beck, 2015). Studies also show that women who have experienced physical or sexual abuse also tend to experience bad health more frequently than other women; and consequently, many become engaged in risky behaviors such as: smoking, physical inactivity, alcohol, and drug abuse (Ellsberg, Jansen, Heise, Watts, & Garcia-Moreno, 2008; Gonzales & Gavillano, 1999; Silverman, Raj, Mucci, & Hathaway, 2001).

## CHALLENGES AND PROBLEMS WITH DOMESTIC VIOLENCE

### Financial

Domestic violence can include financial abuse either during the time the victims are staying with their abusers or after victims/survivors have left their perpetrator. DV victims can be punished through lack of monetary support. Given many women who experience(d) domestic violence tend to be uneducated and unemployed—typically, relegated to household management—they appear to be helpless, according to research studies (Barnett, 2000; Women's Health, 2009). Consequently,

victims/survivors of DV tend to demonstrate a lack of willingness to leave their perpetrators mainly because they lacked autonomy in their financial undertaken (Hauge & Kiamanesh, 2019; Hien & Ruglass, 2008; James, Brody, & Hamilton, 2013; Shilubane & Khoza, 2014). At the same time, those who leave their abusers are faced with the reality that they lack necessary skills to become gainfully employed or earn enough income to survive (ACLU, 2007; Shilubane & Khoza, 2014).

Due to economic abuse and isolation from supportive relationships, the victim usually has little to none of their own savings to rely on and usually few people they can count on to assist them when seeking help (StopVaw, 2010). This has been shown to be one of the greatest obstacles facing victims of DV, and the strongest factor that can discourage them from leaving their perpetrators (StopVAW, 2010). Survey data indicated that 36 percent of US cities reported DV was a primary cause of homelessness in their area (ACLU, 2007). Furthermore, apartment complexes tend to have a *zero-tolerance* policy for crime, to the extent that DV victims may face eviction from their housing and be forced to break their rental agreement (ACLU, 2007). In this situation, the victim is also faced with becoming homeless or having to move to a shelter.

### Personal and Family Safety

According to Hien and Ruglass (2008), victims/survivors of DV “live with intense shame and fear that prevents them from taking action to protect themselves” (p. 50). In addition to shame and fear, several abused women suffer from psychiatric disorders (e.g. PTSD, depression, substance abuse) due to their exposure to violence, which may affect their ability to, “mobilize their psychological and social resources and take action” (Hien & Ruglass, 2008, p. 50). It is well documented that the act of help-seeking can actually be more of a challenge to abused women by slowing down the healing process (Hien & Ruglass, 2008; Leone, Johnson, & Cohan, 2007; Morrison, Luchok, Richter, & Parra-Medina, 2006; WHO, 2016). More specifically, when women face judgmental responses (Morrison et al., 2006), they are left feeling “guilty, depressed, anxious, distrustful of others, and reluctant to seek further help” (Campbell & Raja, 2005, p. 97).

Despite the increasing amount of available shelters, this service is not utilized by most battered women. It was found that less than 2% of severely abused women reported seeking help from a shelter in the past twelve months, while victims of minor violence did not seek help from shelters at all (Clark, Burt, Shulte & Maguire, 1996; Olaniran & Rodriguez, 2013). As a matter of fact, it has been reported that 38% of women murdered globally were victims of IPV (WHO, 2016). Furthermore, because abused women tend to reconcile with their significant other, it is problematic for shelters to help women dealing with DV. As a result, the quality of help for abused women at these shelters is negligible (WHO, 2016). The problem workers at shelters reported to have with abused women returning is due to their measure of success, where success is measured depending on the number of women who do not return without accounting for the reason why they are not returning. There is currently no good evidence supporting the effectiveness of shelters for female victims of DV (Orpin, Papadopoulos, & Puthussery, 2020; Radar Services, 2008).

There is limited research about the effectiveness of hotlines, yet researchers, Dugan, Nagin, and Rosenfeld (2003) noted that “the availability of hotlines in the city, the presence of DV units or training programs in police departments and prosecutors’ offices, and the employment of trained legal advocates on the prosecutor’s staff” each appear to influence retaliation or revictimization by abusive partners (p. 24). Therefore, rather than assisting abuse victims/survivors—as these systems and programs were designed to do—they have been found to backlash (Dugan et al., 2003; WHO, 2016). Unlike traditional resources, DV support/information services via CMC are on the Web in a virtual environment; therefore, these services are not physically present and, perhaps, would not influence revictimization. Another important service for battered women is the health care system, which has not played an active role in identifying or intervening in domestic violence. Research indicates that over one-fifth of women who seek treatment in hospitals display symptoms of DV, however, protocols for DV screening are not invoked or utilized (Orpin et al., 2020; WHO, 2016).

For women who are not connected to other systems, or DV services, it is imperative that hospitals serve as initial intervention point. From there, women may receive information about where to seek help and find other services.

### Technology-Mediated Form of IPV

There are other and non-physical ways for partners or former partners to perpetrate violence against the woman. An emerging and growing trend is the use of communication technology to commit violence against women (Powell & Henry 2018). Using social media, smart phones, email and mobile phones, the abuser can harass a former partner with unwanted phone calls, image sharing, and texting (i.e., hacking). Furthermore, the abuse can be maintained at a constant and high level, without the abuser being physically present. Surveillance apps have been used to enable perpetrators to monitor and stalk either current or former partner (Williams, 2015). New spyware that can be linked to mobile phones is readily available and allows the perpetrator to read partner's email and monitor all other electronic devices. The spyware can identify a victim's current location and can be set up to mirror the phone or a personal computer where the perpetrator can watch everything the victim is doing. Seldomly, a tracking device may be attached to the woman's car so that she can be easily stalked (Stolz, 2017). Indeed, the woman or victim may not even be aware of the illegal tracking until the device is located by auto mechanics. Technology-mediated violence against women can traumatize and isolate women, because online interactions relate to daily part of life. The Internet and mobile phones are used to maintain social contact, but due to tech stalking, women may have to change their online persona or withdraw their digital presence to stop their abuser (Powell & Henry 2018). Many of the devices used by perpetrators to intimidate and stalk are easily available, and improvements in technology are likely to make them even more so. Consequently, health care service providers need to be included in how to recognize potential these technological abuses.

### Domestic Violence Against Migrant Women Workers

There are two types of violence when dealing with migrant women workers (MWW). First, there is violence against MWW during employment and the other is domestic violence against MWWs. However, these two categories are not mutually exclusive. Immigrants, in general face some challenges when they resettle in a foreign country. A host of factors influence their experience, including the resources they bring to the host country and those they find there. Immigrant women find themselves unable to participate or navigate the same resource networks as their male cohorts (Huysman, 2014; Olaniran, 2018). In some instances, men serve as intermediaries between the women and community/state resources. Thus, women are unable to navigate services or available resources on their own, their male partners may be the determinant of what services women are able to access in terms of resources (Menjívar & Salcido, 2002; Olaniran, 2018).

Furthermore, *isolation* has been suggested to exacerbate violence against MWW in the sense that it is easier for men or intimate partners to control these women both emotionally and physically in this manner (Sabina, Cuevas, & Zadnik, 2014). Because of isolation, intimate partners can gain control over resources that could offer legal, financial, and/or emotional support to these women (Hauge & Kiamanesh, 2019). For instance, it was discovered that when Guatemalan and Salvadoran women in the United States received information on domestic violence and their rights at community organizations, their partners were not appreciative of such knowledge (Menjívar, 2000).

From a different standpoint, culture (i.e., native culture and host culture) impacts MWW's predisposition to intimate partner violence (IPV). It has been shown that Latino women's experiences of IPV are drastically impacted by their culture and cultural adaptation to the United States (Grzywacz, et al., 2009; Orpin, et al., 2017; Sabina, et al., 2014, Yoshihama, Blazevski, & Bybee, 2019). Specifically, MWWs' adaptation to Anglo orientation in USA is presumed to change the family dynamics of the Latinos in terms of redistribution of power which often causes tension (Grzywacz, et al., 2009) when individuals attempt to renegotiate their roles in their new environment. Intimate partners may view

the new cultural shifts in Latino women as threatening to the woman's traditional roles in the family. This consequently increases the risk of IPV (Sabina et al., 2015).

### **CMC and IPV Victims/Survivors**

For the most part, CMCs role in IPV is that of information dissemination. That is, CMC offers the ability to link people together in a manner where they can discuss ideas and share advocacy strategies in a quick and cost-effective way. These interactions and discussions are facilitated by social media, e-mail, listservs, telnet, and teleconferencing in gaining insights to violence against women. At the same time, the information dissemination that occurs via CMC has made the issue of DV a global, rather than local, one by addressing violence against women irrespective of age, class, race, and ethnicity (Tafnout & Timjerdine, 2009).

The Internet has also provided DV organizations with a greater flexibility to reach out to victims/survivors of domestic violence than previously possible through print and word-of-mouth outreach efforts, because it can reach a mass audience (Finn, 2000; Olaniran, 2018). CMC offers new approaches for outreach and a new arena for service delivery, which is very attractive to a movement committed to social change. While there are many advantages, DV organizations and caregivers are cautioned to temper enthusiasm with critical assessment of CMC. The digital divide continues to be an important issue in CMC usage and adoption in both developed and developing worlds (Olaniran, 2007, 2010; Zheng, 2016). Recent studies confirm existence of the digital divide and found significant correlations among variables associated with differences in economic development levels. For instance, countries with higher levels of CMC adoption showed positive and significant correlation with gross domestic product (GDP), service sector, education, and governmental effectiveness. However, in developing countries, population, age, and urban population are positively associated with CMC adoption, while Internet costs impact access and usage negatively (e.g., Billon, Marco, & Lera-Lopez, 2009; Zheng, 2016). Consequently, caregivers and domestic violence organizations are encouraged to address issues surrounding the digital divide to understand who benefits from their services—especially those offered via CMC and technology platforms. These service organizations also need to weigh the financial costs of service delivery with budget allocation decisions before embarking on service delivery through CMC. DV organizations must ensure that CMC usage and technologies deployed do not further ostracize parts of their targeted audience or potential users like those who do not have access to the technologies and disabled those who may very well need their provided services (Olaniran, 2010).

Depending on the type of ICT support used, there is the issue of continue usage beyond initial adoption that caregivers and DV organizations need to bear in mind. A study in e-health indicated that e-mail support interventions benefit some but not all caregivers and these interventions have high non-usage attrition (Chiu et al., 2009). Using the Unified Theory of Acceptance and Use of Technology (UTAUT), which explains the intention to use information technology (Venkatesh, Morris, Davis, & Davis, 2003), there are four principles including: (1) Performance expectancy, which deals with a person's belief that a new technology will improve task goals; (2) Effort expectancy, which is the degree of ease with the use of a new technology. This is similar to the Ease of Use (EOU) addressed by Olaniran (2007); (3) Social influence, which addresses the level of influence other people exert in their decision to use new technology; and, (4) Facilitating conditions focus on the level of support a user has available in using a given technology. These four principles affect intention to use and in predicting technology use. As a matter of fact, Venkatesh et al. (2003) found that performance expectancy and effort expectancy explain significant proportion of the variance than any of the other factors.

In the utilization stage attrition or discontinuation by caregivers occurs. Based on factors such as clinical needs and technology aptitude, caregivers with less competence were less likely to continue service and those with a more positive attitude toward technology were more likely to continue service. Furthermore, upon identifying and experiencing clinical benefits, frequent system users were more likely to continue service and concluded that they benefited more from CMC. Consequently, Chiu and

Eysenbach (2010) concluded that while usage behaviors are influenced primarily by technological factors in the early stages of adoption both clinical and technological factors are critical in the later stages of adoption; to the extent that, the frequency of use is influenced by clinical outcomes.

There are other social challenges or barriers that impact the utilization of ICTs in care giving for DV victims/survivors. Olaniran (2007) alluded to general social challenges including digital divide, illiteracy, and technology illiteracy among others. At times there are those who have access in their own homes but are afraid that their abusers may be monitoring their usage or ultimately prevent them from using CMCs. Given the socio-economic status of women, access to technology is not immediately a given. Some could not afford computer or Internet service, and some have to contend with choosing between access (e.g., mobile phone with Internet access) and putting food on the table thus, they may have to rely on public assistance. It is alarming that this is the case with domestic violence victims/survivors regardless of age, education and socio-economic class. For instance when one talks about digital divide in CMC use, traditionally one thinks of the haves and have-nots, which is often compounded by the fact that individuals from Economically Developed Countries (EDCs) usually enjoy more readily available access to CMCs than those from Less Economically Developed Countries (LEDCs) (Olaniran, 2007). Whereas, many individuals (e.g., less educated, poor, and other marginalized groups) will be left out. Hence, the issue of technical literacy persists, which hinders individuals from the usage of such technology (Olaniran, 2007).

Although the Internet provides a nice medium and several avenues for seeking information and services for DV victims/survivors, the Web itself also creates a means for victimization—particularly when it comes to privacy. There is the danger of loss of privacy for individuals who access chat rooms and post messages on discussion boards and forums (Briggs, 2018; Finn, 2000; Olaniran & Rodriguez, 2013). This is because messages tend to be archived and can be searched by abusers, who then have access to the location of their victims and ultimately cyber-stalk and harass them (Briggs, 2018; Kranz, 2002). Also, while CMCs are instrumental in raising awareness and providing necessary coping mechanisms to DV victims/survivors, they can also become a tool that abusers use to control their victims. For example, for victims who have or possess technological means, such as cell phone or Internet access, they may have no control over them because they can be deprived of them at any time or when their partners get angry. Tafnout and Timjerdine (2009) offer the plight of DV victims in Morocco, where one victim contends: “I have to tell my husband because he must agree. The man sees the cell phone as an enemy. If the man has the right to have the cell phone, the woman also has the right to have it and needs it” (p. 94). Thus, female victims/survivors may not feel that they are able to benefit fully or use ICTs to cope with their DV experiences.

## **TOWARDS AI, IOT, CLOUD COMPUTING AS SOLUTIONS**

For ICTs to succeed as agents of change in the lives of DV victims/survivors, certain things must be in place. Caregivers and counselors must help empower women to be self-confident, along with trusting in their own abilities to bring about the desired change. For instance, victims must be able to rely on their own power and recognize the need to participate or become proactive in decision makings about turning unpleasant situations around, despite social, economic, physiological, and family impacts. They must conclude or realize that they can gain control of their lives and fate, and the lives and destiny of their families and children. More importantly, victims of DV/IPV should be helped in a manner where, they not only know how to move from a state of helplessness to confidence in their own abilities, but also to be empowered to become advocates for other victims/survivors using CMCs. Tafnout and Timjerdine (2009) points to how legal aid shelters for DV victims in Morocco was able to accomplish this goal, as women who possessed CMC tools were made aware of the strategic role of CMC in protecting and in opening up to what is happening outside their surroundings. According to the women, the mobile phone replaces the role of the family in reporting. As one IPV female survivor puts it: “In the past when your husband beats you, you seek help from your family who asks you to

be patient, but now you can contact a legal center which can make your problem known and thus you help other women to reject violence” (Tafnout & Timjerdine, 2009, p. 93). Similarly, by encouraging the use of ICTs for support and information services, female victims/survivors can access, retrieve, and develop their own understanding based upon their personal experiences and needs.

Another recommendation is to ensure that all DV service providers (e.g., shelters, caregivers, hospitals, clinics, counselors, libraries, and so on) are equipped to provide Internet access, as well as ensuring that their staff is computer literate. Similarly, education to community groups and businesses can include Internet access to domestic violence services as part of the training. Trainers can encourage other organizations to bookmark links to domestic violence services on their computers and mobile phones, can offer links of local and national domestic violence resources (Olaniran & Rodriguez, 2013). Alternatively, there is the suggestion to encourage and promote Internet access in public places (e.g., libraries and community centers, where it may be safer for individuals to access DV resources that are available online and more importantly, where abusers cannot track victims/survivors’ browsing history.

In addressing domestic violence, caregivers are susceptible to vicarious trauma. Therefore, the best way for a clinician or caregiver to help avoid developing vicarious trauma is to incorporate good self-care practices. These can include exercise and relaxation techniques, debriefing with colleagues, and seeking support from supervisors (Olaniran & Rodriguez, 2013). Additionally, it is recommended that clinicians make the positive and rewarding aspects of working with domestic violence victims/survivors the primary focus of thought and energy, such as being part of the healing process. Clinicians should also continually evaluate their empathetic responses to victims, in order to avoid being sucked into the trauma that the victims are experiencing. Subsequently, it is recommended that clinicians observe good boundaries, and find a balance in expressing empathetic responses to the victim, while still maintaining personal separation from their clients’ traumatic experiences.

Although, ICTs are used to perpetuate IPV as discussed earlier, they can also be used to provide plethora of opportunities to reduce IPV. For example, the interconnectedness of technological devices through wearables otherwise known as the Internet of Things (IoT) and data mining through big data offer potential in this regard. Of importance is the notion body area network (BAN) or monitored features that collect and gather data, process it, and identify useful information (Yuce, 2010). BAN involves all of the applications and communication on, in, and near the body for monitoring physiological signals to detect a reaction to an attack (Rodriguez-Rodriguez, Rodriguez, Moreno, Heras-Gonzalez, & Gentili, 2019). At present, smartphones offer 24/7 monitoring capability with their accelerometers and GPS that enable to a mechanism for managing information. Furthermore, they can send data and information to the cloud, to be either stored, or forward emergency calls in case a survivor is at risk.

Similarly, there are other commercial devices, such as the Amazon Echo, which is equipped with an artificial intelligence (AI), like Apple Siri (Apple), and Google Home that allows monitoring beyond recreational usage (Rodriguez-Rodriguez, et al., 2019). There are also applications in the field of home security, such as the commercial Alexa Guard, which provides an alarm when noises are detected (such as breaking glass), and can also detect a cry for help that can be adapted to the management and prevention of IPV (Huang, Chiew, Li, Kok, & Biswas, 2015; Rodriguez-Rodriguez, et al., 2019). It needs to be said that that these technologies are not necessarily fool-proofed because they can be hacked or used by IPV offenders as well. However, the technologies still offer another layer of protection for victims of IPV.

## **FUTURE DIRECTION**

In the future, it would be interesting to see which types of ICT mediums are preferred or used the most by IPV victims/survivors (e.g., chat rooms, searching for information, online counseling, etc.).



Future research can assess and evaluate which applications or learning tools are preferred by DV victims/survivors in general.

Another direction for the future is to examine the impact of new technology infrastructure or platform such as the 5Gs and the ability to deliver services in locales that are still dealing with digital divide or lack of access to the state of the art mobile phones and wireless technologies. One area of DV that is under explored is that of male victims/survivors of DV. It is troubling that there is not much literature in this area when in fact there are vivid evidence that males also suffer from domestic violence and IPV. The Bureau of Justice statistics indicated that of the 2,340 DV deaths in 2007, 30% were male. Perhaps the lack of literature on male as victims of DV is because the information is rarely reported by men because of the taboo surrounding men as victims of IPV especially when the perpetrators of such violence are women. More importantly, help providers and healthcare givers must be aware that ICT usage is not a panacea for IPV victims/survivors notwithstanding.

## CONCLUSION

The use of information and communication technology, specifically CMC—will continue to increase and expand across all boundaries and collectives of individuals—users, organizations, communities, businesses, and so on. Researchers (Finn, 2000, Kranz, 2002, Olaniran, 2007) continuously examine the nature of use and effectiveness of ICTs, as technology rapidly changes the ways individuals and groups interact and provide information/support services, among others. This paper examines the nature of IPV, the role that ICTs play in providing support services to victims/survivors of DV and in particular IPV, while offering some recommendations. Finally, as ICTs (e.g., social media's, networks, websites) are gaining popularity even as IPV continues to be a major issue globally. Therefore, the role of Internet (i.e., IoTs, cloud computing, and AI) may be in infancy in terms of how to deploy ICTs to assist in IPV.

## REFERENCES

- American Civil Liberties Union Women's Rights Project (ACLU). (2007). Retrieved April 24, 2010, from <https://www.aclu.org/FilesPDFs/housing%20paper.4.pdf>
- Appel, A. E., & Holden, G. W. (1998). The co-occurrence of spouse and physical child abuse: A review and appraisal. *Journal of Family Psychology*, 12(4), 578–599. doi:10.1037/0893-3200.12.4.578
- Barnett, O. W. (2000). Why battered women do not leave, part 1: External inhibiting factors within society. *Trauma, Violence & Abuse*, 1(4), 343–372. doi:10.1177/152483800001004003
- Bergen, R. K. (1996). *Wife rape: understanding the response of survivors and service providers*. Sage. doi:10.4135/9781483327624
- Billon, M., Marco, R., & Lera-Lopez, F. (2009). Disparities in ICT adoption: A multidimensional approach to study the cross-country digital divide. *Telecommunications Policy*, 33(10-11), 596–610. doi:10.1016/j.telpol.2009.08.006
- Breiding, M. J., Black, M. C., & Ryan, G. W. (2008). Chronic disease and health risk behaviors associated with intimate partner violence—18 U.S. states/territories, 2005. *Annals of Epidemiology*, 18(7), 538–544. doi:10.1016/j.annepidem.2008.02.005 PMID:18495490
- Brennan, P. F., Moore, S. M., & Smyth, K. A. (1992). Alzheimer's disease caregivers' uses of a computer network. *Western Journal of Nursing Research*, 14(5), 662–673. doi:10.1177/019394599201400508 PMID:1529609
- Briggs, C. (2018). An emerging trend in domestic violence: Technology-facilitated abuse. *Australian Journal of Child & Family Health Nursing*, 15(15), 1–2.
- Bureau of Justice Statistics. (n.d.). *Intimate partner violence*. Available from URL: <http://bjs.ojp.usdoj.gov/index.cfm?ty=tp&tid=971#summary>
- Campbell, K., Sy, S., & Anderson, K. (2000). On-line learning for abused women and service providers in shelters: Issue of representation and design. *Canadian Journal of University Continuing Education*, 26(1), 15–51.
- Campbell, R., & Raja, S. (2005). The sexual assault and secondary victimization of female veterans: Help-seeking experiences in military and civilian social systems. *Psychology of Women Quarterly*, 29(1), 97–106. doi:10.1111/j.1471-6402.2005.00171.x
- Centers for Disease Control and Prevention (CDC). (2003). *Costs of intimate partner violence against women in the United States*. Atlanta, GA: CDC, National Center for Injury Prevention and Control. Retrieved February 22, 2011, from [www.cdc.gov/ncipc/pub-res/ipv\\_cost/ipv.htm](http://www.cdc.gov/ncipc/pub-res/ipv_cost/ipv.htm)
- Centers for Disease Control and Prevention (CDC). (2017). *National Intimate Partner and Sexual Violence Survey (NISVS)*. [https://www.cdc.gov/ViolencePrevention/pdf/NISVS\\_FactSheet-a.pdf](https://www.cdc.gov/ViolencePrevention/pdf/NISVS_FactSheet-a.pdf)
- Chiu, T., & Eysenbach, G. (2010). Stages of use: Consideration, initiation, utilization, and outcomes of an internet-mediated intervention. *BMC Medical Informatics and Decision Making*, 10(1), 73. doi:10.1186/1472-6947-10-73 PMID:21092275
- Chiu, T., Marziali, E., Colantoni, A., Carswell, A., Gruneir, M., Tang, M., & Eysenbach, G. (2009). Internet-based caregiver support for Chinese Canadians taking care of a family member with Alzheimer disease and related dementia. *Canadian Journal on Aging*, 28(4), 323–336. doi:10.1017/S0714980809990158 PMID:19925698
- Clark, S. J., Burt, M. R., Schulte, M. M., & Maguire, K. (1996, October). *Coordinated community responses to domestic violence in six communities: Beyond the justice system (summary)*. Special report submitted to U.S. Department of Health and Human Services. Washington, DC: The Urban Institute.
- Coker, A. L. (2006, March). Preventing intimate partner violence: How we will rise to this challenge. *American Journal of Preventive Medicine*, 30(6), 528–529. doi:10.1016/j.amepre.2006.03.002 PMID:16704948
- Crofford, L. J. (2007). Violence, stress, and somatic syndromes. *Trauma, Violence & Abuse*, 8(3), 299–313. doi:10.1177/1524838007303196 PMID:17596347

- Demaris, A., Benson, M. L., Fox, G. L., Hill, T., & Van Wyk, J. (2003). Distal and proximal factors in domestic violence: A test of an integrated model. *Journal of Marriage and Family*, 65(3), 652–667. doi:10.1111/j.1741-3737.2003.00652.x
- Dugan, L., Nagin, D. S., & Rosenfeld, R. (2003). Do domestic violence services save lives? *NIJ Journal*, 1(250), 20–25.
- Ehrensaft, M. K. (2007, October). Intimate partner violence: Persistence of myths and implications for intervention. *Children and Youth Services Review*, 30(3), 276–286. doi:10.1016/j.childyouth.2007.10.005
- El-Morr, C., & Layal, M. (2019, March). ICT-Based Interventions for Women Experiencing Intimate Partner Violence: Research Needs in Usability and Mental Health. In ITCH (pp. 103-109). Academic Press.
- Ellsberg, M., Jansen, H. A., Heise, L., Watts, C. H., & Garcia-Moreno, C. (2008). Intimate partner violence and women's physical and mental health in the WHO multi-country study on women's health and domestic violence: An observational study. *Lancet*, 371(9619), 1165–1172. doi:10.1016/S0140-6736(08)60522-X PMID:18395577
- Finn, J. (2000). Domestic violence organizations on the web: A new arena for domestic violence services. *Violence Against Women*, 6(1), 80–102. doi:10.1177/10778010022181723
- Gelles, R. J., & Straus, M. A. (1988). *Intimate Violence*. Simon and Schuster.
- Gondolf, E. W. (1998). *Assessing women battering in mental health services*. Sage. doi:10.4135/9781483328188
- Gondolf, E. W., & Fisher, E. R. (1988). *Battered women as survivors: An alternative to treating learned helplessness*. D. C. Health.
- Gonzales, D. E., & Gavillano, L. (1999). Does poverty cause domestic violence? Some answers from Lima. In A. Morrison & M. Biehl (Eds.), *Too close to home: Domestic violence in the Americas*. (pp. 35-49). Washington, DC: Inter-American Development Bank.
- Grzywacz, J. G., Rao, P., Gentry, A., Marín, A., & Arcury, T. A. (2009). Acculturation and conflict in Mexican immigrants' intimate partnerships: The role of women's labor force participation. *Violence Against Women*, 15(10), 1194–1212. doi:10.1177/1077801209345144 PMID:19706779
- Hamm, S. (2001). Information communications technologies and violence against women. *The Society for International Development*, 44(3), 36–41. doi:10.1057/palgrave.development.1110259
- Hauge, M., & Kiamanesh, P. (2019). Mothering and everyday life during and in the aftermath of domestic violence among women with immigrant background in Norway. *Child & Family Social Work*, 1–8. doi:10.1111/cfs.12710
- Hick, S., Halpin, E., Hoskins, E., Robinson, M., & Hussein, A. (Eds.). (2000). *Human Rights and the Internet*. Macmillan. doi:10.1057/9780333977705
- Hien, D. A., & Ruglass, L. M. (2008, November). Interpersonal partner violence and women in the United States: An overview of prevalence rates, psychiatric correlates and consequences, and barriers to help seeking. *International Journal of Law and Psychiatry*, 32(1), 48–55. doi:10.1016/j.ijlp.2008.11.003 PMID:19101036
- Huang, W., Chiew, T. K., Li, H., Kok, T. S., & Biswas, J. (2010). Scream detection for home applications. *Proceedings of the 2010 5th IEEE Conference on Industrial Electronics and Applications*, 2115–2120. doi:10.1109/ICIEA.2010.5515397
- Huysman, M. (2014). Knowledge sharing, communities, and social capital. *Communities of Practice*, 77.
- James, L., Brody, D., & Hamilton, Z. (2013). Risk factors for domestic violence during pregnancy: A meta-analytic review. *Violence and Victims*, 28(3), 359–380. doi:10.1891/0886-6708.VV-D-12-00034 PMID:23862304
- Johnson, M. P., & Ferraro, K. J. (2000). Research on domestic violence in the 1990s: Making distinctions. *Journal of Marriage and Family*, 62(4), 948–963. doi:10.1111/j.1741-3737.2000.00948.x
- Kranz, A. L. (2002, May). Helpful or harmful? How innovative communication technology affects survivors of intimate partner violence. *MINCAVA Electronic Clearing House*. Retrieved September 18, 2010 from, <http://www.mincava.umn.edu/documents/5survivortech/5survivortech.html>

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2 • July-December 2021

Kranz, A. L., & Nakamura, K. (2002). *Helpful or harmful: How innovative communication technology affects survivors of intimate partner violence*. Minneapolis: University of Minnesota, Minnesota Center Against Violence and Abuse. Retrieved March 24, 2011, from <http://www.mincava.umn.edu>

Leone, J. M., Johnson, M. P., & Cohan, C. L. (2007, December). Victim help seeking: Differences between intimate terrorism and situational couple violence. *Family Relations*, 56(5), 427–439. doi:10.1111/j.1741-3729.2007.00471.x

Leserman, J., & Drossman, D. A. (2007). Relationship of abuse history to functional gastrointestinal disorders and symptoms. *Trauma, Violence & Abuse*, 8(3), 331–343. doi:10.1177/1524838007303240 PMID:17596349

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLoS One*, 13(8), e0199661. doi:10.1371/journal.pone.0199661 PMID:30067747

Mahoney, D. F., Tarlow, B., Jones, R. N., Tennstedt, S., & Kasten, L. (2001). Factors affecting the use of a telephone-based intervention for caregivers of people with Alzheimer's disease. *Journal of Telemedicine and Telecare*, 7(3), 139–148. doi:10.1258/1357633011936291 PMID:11346473

Max, W., Rice, D. P., Finkelstein, E., Bardwell, R. A., & Leadbetter, S. (2004). The economic toll of intimate partner violence against women in the United States. *Violence and Victims*, 19(3), 259–272. doi:10.1891/vivi.19.3.259.65767 PMID:15631280

McCauley, J., Kern, D. E., Kolodner, K., Dill, L., Schroeder, A. F., DeChant, H. K., Ryden, J., Bass, E. B., & Derogatis, L. R. (1995). The battering syndrome prevalence and clinical characteristics of domestic violence in primary health care internal medicine practices. *Annals of Internal Medicine*, 123(10), 737–746. doi:10.7326/0003-4819-123-10-199511150-00001 PMID:7574191

Menjívar, C. (2000). *Fragmented ties: Salvadoran immigrant networks in America*. University of California Press.

Menjívar, C., & Salcido, O. (2002). Immigrant women and domestic violence: Common experiences in different countries. *Gender & Society*, 16(6), 898–920. doi:10.1177/089124302237894

Morrison, K. E., Luchok, K. J., Richter, D. L., & Parra-Medina, D. (2006). Factors influencing help-seeking from informal networks among African American victims of intimate partner violence. *Journal of Interpersonal Violence*, 21(11), 1493–1511. doi:10.1177/0886260506293484 PMID:17057164

National Center for Injury Prevention and Control (NCIPC). (2003). *Costs of intimate partner violence against women in the United States*. Centers for Disease Control and Prevention.

Olaniran, B. A. (1994). Group performance and computer-mediated communication. *Management Communication Quarterly*, 7(3), 256–281. doi:10.1177/0893318994007003002

Olaniran, B. A. (1995). Perceived communication outcomes in computer-mediated communication: An analysis of three systems among new users. *Information Processing & Management*, 31(4), 525–541. doi:10.1016/0306-4573(95)00006-3

Olaniran, B. A. (2007). Challenges to implementing e-learning in lesser developed countries. In A. L. Edmundson (Ed.), *Globalized e-learning cultural challenges*, (pp. 18–34). Hershey, PA: Idea Group, Inc. doi:10.4018/978-1-59904-301-2.ch002

Olaniran, B. A. (2010). Challenges Facing the Semantic Web and Social Software as Communication Technology Agents in E-learning Environments. *International Journal of Virtual and Personal Learning Environments*, 1(4), 18–30. doi:10.4018/jvple.2010100102

Olaniran, B. A. (2018). Violence against migrant women workers (MWWs). In Y. Mao & R. Ahmed (Eds.), *Culture, Migration and Health Communication in a Global Context* (pp. 156–174). Routledge.

Orpin, J., Papadopoulos, C., & Puthussery, S. (2020). The prevalence of domestic violence among pregnant women in Nigeria: A systematic review. *Trauma, Violence & Abuse*, 21(1), 3–15. doi:10.1177/1524838017731570 PMID:29333978

Powell, A., & Henry, N. (2018). *Digital harassment and abuse of adult Australians: a summary report*. RMIT University.

Radar Services, Inc. (2008). *Why have domestic violence programs failed to stop abuse?* RADAR Services, Inc.

- Roberts, T. A., Klein, J. D., & Fisher, S. (2003). Longitudinal effect of intimate partner abuse on high-risk behavior among adolescents. *Archives of Pediatrics & Adolescent Medicine*, 157(9), 875–981. doi:10.1001/archpedi.157.9.875 PMID:12963592
- Rodríguez-Rodríguez, I., Rodríguez, J. V., Elizondo-Moreno, A., Heras-González, P., & Gentili, M. (2020). Towards a Holistic ICT Platform for Protecting Intimate Partner Violence Survivors Based on the IoT Paradigm. *Symmetry*, 12(1), 37. doi:10.3390/sym12010037
- Sabina, C., Cuevas, C., & Zadnik, E. (2015). Intimate Partner Violence among Latino Women: Rates and Cultural Correlates. *Journal of Family Violence*, 30(1), 35–47. doi:10.1007/s10896-014-9652-z
- Saltzman, L. E., Fanslow, J. L., McMahon, P. M., & Shelley, G. A. (2002). *Intimate partner violence surveillance: Uniform definitions and recommended data elements, version 1.0*. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Available from [https://www.cdc.gov/ncipc/pub-res/ipv\\_surveillance/intimate.htm](https://www.cdc.gov/ncipc/pub-res/ipv_surveillance/intimate.htm)
- Shilubane, H. N., & Khoza, L. B. (2014). Violence against women in Limpopo Province, South Africa: Women's health issues. *African Journal for Physical Health Education, Recreation and Dance*, 20, 83–93.
- Silverman, J. G., Raj, A., Mucci, L., & Hathaway, J. (2001). Dating violence against adolescent girls and associated substance use, unhealthy weight control, sexual risk behavior, pregnancy, and suicidality. *Journal of the American Medical Association*, 286(5), 572–579. doi:10.1001/jama.286.5.572 PMID:11476659
- Simmons, C. A., Lindsey, L., Delaney, M. J., Whalley, A., & Beck, J. G. (2015). Real-world barriers to assessing and treating mental health problems with IPV survivors: A qualitative study. *Journal of Interpersonal Violence*, 30(12), 2067–2086. doi:10.1177/0886260514552275 PMID:25304669
- Stolz, G. (2017). Disturbing new trend: domestic violence offenders use car tracking. *Sunday Mail*. <https://www.couriermail.com.au/news/queensland/crime-and-justice/disturbing-new-trend-domestic-violence-offenders-use-car-tracking/news-story/b5bf4cbadb57fc91998e7b3ea7d5bb4>
- Stop Violence Against Women (StopVAW). (2010). Retrieved April 24, 2010, from [https://www.stopvaw.org/Domestic\\_Violence\\_and\\_Housing.html](https://www.stopvaw.org/Domestic_Violence_and_Housing.html)
- Tafnout, A., & Timjerdine, A. (2009). Using ICT to act on hope and commitment: The fight against gender violence in Morocco. In I. Buskens & A. Webb (Eds.), *African Women and ICTs: Investigating Technology, Gender, & Empowerment* (pp. 88-96). New York: Zed Books.
- Tjaden, P., & Thoennes, N. (1998). *Stalking in America: Findings from the National Violence Against Women Survey*. Washington, DC: Department of Justice (US); Publication No. NCJ 169592. Available from: <https://www.ncjrs.gov/pdffiles/169592.pdf>
- Tjaden, P., & Thoennes, N. (2000). *Extent, nature, and consequences of intimate partner violence: Findings from the National Violence Against Women Survey*. Washington, DC: Department of Justice (US). Publication No. NCJ 181867. Retrieved January 31, 2011, from [www.ojp.usdoj.gov/nij/pubs-sum/181867.htm](http://www.ojp.usdoj.gov/nij/pubs-sum/181867.htm)
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. (2003). User acceptance of information technology: Toward a unified view. *Management Information Systems Quarterly*, 27(3), 425–478. doi:10.2307/30036540
- Walker, L. (1979). *The battered woman*. Harper and Row.
- Williams, R. (2015). Spyware and smartphones: how abusive men track their partners. *The Guardian*. <https://www.theguardian.com/lifeandstyle/2015/jan/25/spyware-smartphoneabusive-men-track-partners-domestic-violence>
- Women's Health Organization. (2009, January). *Domestic and intimate partner violence*. Retrieved August 4, 2009, from <https://www.womenshealth.gov/violence/types/domestic.cfm>
- Woodlock, D. (2016). The abuse of technology in domestic violence and stalking. *Violence Against Women*, 23(5), 584–602. doi:10.1177/1077801216646277 PMID:27178564
- World Health Organization. (2016). *Media center: Violence against women*. Retrieved from [http://www.who.int/mediacentre/fact\\_sheets/fs239/en/](http://www.who.int/mediacentre/fact_sheets/fs239/en/)

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2 • July-December 2021

Yoshihama, M., Blazeovski, J., & Bybee, D. (2019). Gender (a) symmetry in correlates of perpetration of intimate partner violence: Gender (a) symmetry in IPV and the role of gender attitudes. *Violence Against Women*. PMID:31187698

Yuce, M. R. (2010). Implementation of wireless body area networks for healthcare systems. *Sensors and Actuators. A, Physical*, 162(1), 116–129. doi:10.1016/j.sna.2010.06.004

Zheng, H. (2016). A study on the usability of e-commerce websites between China and Thailand. *International Journal of Simulation: Systems, Science and Technology*, 17(1), 34–51.

*Bolanle A. Olaniran is a professor in the communication Studies department in the college of Media and Communication at Texas Tech University, Lubbock, TX, USA.*

# Using Data Science Software to Address Health Disparities

Jose O. Huerta, University of North Texas, USA

Gayle L. Prybutok, University of North Texas, USA

Victor Prybutok, University of North Texas, USA

## ABSTRACT

The article assesses data science software to evaluate the usefulness of data science technology in addressing concerns such as health disparities. Data science software was analyzed using KDnuggets data related to analytics, data science, and machine learning software. Data science functionalities include computational processes and frameworks that are relevant for healthcare. This study demonstrates the importance of leading applications for conducting data science operations that can improve care in healthcare networks by addressing such factors as health disparities.

## KEYWORDS

Analytics Software, Data Science, Data Science Software, Health Analytics Software, Health Disparities

## INTRODUCTION

The application of data science in health care was studied by many professionals in the health care space to forecast its value and particular uses. Although data science is a beneficial tool for new knowledge and insights in healthcare, there exist challenges to its application in the domain. These challenges include data accuracy, missing data, and standardizing of data (Delaney & Westra, 2016). Although these are very important challenges to address, an important axiom to keep in mind is that the underlying information complexity to be achieved would have a major effect on the information system structure most appropriate for achieving the desired information outcome (Murphy, Murphy, Buettner, & Gill, 2015). In addition to these challenges, healthcare specialties such as the biomedical field have had challenges acquiring, sharing, and analyzing data (Dunn & Bourne, 2017). Therefore, data science in healthcare may in some ways be limited, but it is nonetheless useful to help solve significant and common healthcare problems. One such problem is that of health disparities found across health care organizations. Addressing health disparity issues allow health organizations to optimize patient care approaches and improve outcomes. Health care organizations can benefit through the impact data science software can have on their organizations and the multiple ways data science can lead to important findings in health care. For instance, the Covid-19 pandemic media coverage has reported mortalities among blacks in the United States at a higher rate compared to Caucasians (Shelby Lin Erdman, 2020). Is this due to disparities in socioeconomic issues and healthcare access that

DOI: 10.4018/IJBDAH.20210701.oa4

This article, published as an Open Access article on April 23, 2021 in the gold Open Access journal, International Journal of Big Data and Analytics in Healthcare (converted to gold Open Access January 1, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

ultimately may lead to the mortality rate? While not the focus of the paper, it is important to recognize the potential that is driving the development and examination throughout the paper and where data science can offer some promise. The present study will review data science in relation to addressing health disparities in healthcare. It assesses data science software to examine the effectiveness of data science technologies that may be used to address problems such as health disparities.

## LITERATURE REVIEW

Applications of data science are evident in numerous fields, ranging from research-based disciplines such as market, social, and census research to financial, technical, consulting, business, and media disciplines (Fayyad, 2012). The field of healthcare has begun to benefit from data science amid acquisition of new healthcare technologies. These new technologies also make available new opportunities for data science exploration, which can lead to intriguing discoveries from the data collected. For example, data science can be an important component of health informatics. Although viewed with some skepticism initially, health informatics has been embraced by the healthcare industry over time through vital investments in health information technology (HIT), increasing exploration of its utility (Detmer & Shortliffe, 2014). Data science may have similar adoption challenges, but as data begins to increase at a rapid rate, embracing data science as a discipline and new technology will soon begin to make sense. For example, data science software can be important to clinicians because it can reduce unnecessary expenses in patient care, improve care quality and patient safety, and streamline the patient care process. Additionally, data science can help to determine the level of care or the level of care transitions that must occur for the well-being of the patient. Such information can come from new insights surfaced in the application of data science to patients' health improvement.

### Data Production

Data science has come to cohere as a recognized field internationally, crossing numerous disciplines over decades, and evolving to respond to new data technologies (Liu et al., 2009; Press, 2013; Smith, 2006). As the healthcare field has met new data challenges in recent years, data science has offered powerful tools. It is of high value to note that big data and its application in the healthcare industry help to cut costs from analysis performed from electronic medical records (De la Torre Diez, Cosgaya, Garcia-Zapirain, & Lopez-Coronado, 2016). Such benefits have been made possible by innovations in managing large data sets. The importance of digital data for science is growing, and methods for analyzing these data need new data analytics (Westra, 2017).

The field of healthcare is witnessing an ever-increasing generation of large and complex data sets, commonly called *big data*, a term that functions as a shorthand for the diverse objects of data science (Rumbold & Pierscionek, 2017). Healthcare has experienced big data increase and therefore makes data science approaches to information promising. For example, in GIS applications, big data can boost monitoring of public health by combining spatial variables and social health determinants (Zhang et al., 2017). More specifically, Allen, Tsou, Aslam, Nagel, & Gawron (2016) conducted a study that utilized geographical information systems (GIS) methodologies using data mined from social media platforms, leveraging techniques in machine learning, a component of data science, to filter through the data before analysis. Data science features a broad variety of techniques including mining text, visualizing data, geospatial modeling, machine learning, and predictive analysis. (O'Connor, 2018). Health care has seen the advancement of data science due to the following: big data, new data produced from sources that emerge from clinical trials and research, and the new technological capacities available for creating and deciphering data, whether structured or unstructured (Baptista et al., 2019). The industry of healthcare is positioning itself to retrieve valuable insights from data science technologies and processes, which help to produce noteworthy value, aiding in the significant utilization of data science methods and data science software for health care applications.



For example, information from electronic health records and other organizations such as the Center for Medicare and Medicaid Services (CMS) produce clinical data sets that allow for its use across multiple important settings in health care (Chase & Vega, 2016). Data sets can store information particular to the population, such as demographics, which can help aid in research when incorporating other factors such as income into the study. In turn, this can help researchers highlight gaps based on the subject matter content of the study (Chase & Vega, 2016). These types of health-centric data are necessary for healthcare data science applications, and there can be a significant improvement in the analysis of data. Organizations such as CMS can benefit from finding relevant and deep insights buried among the complexity of variables and attributes that can exist in their data. Other healthcare organizations that work closely with CMS do so through multi-disciplinary aspects that exist in many forms, such as that of finance, management, and even policy; especially policy that can have a major impact on many health care disciplines that must adhere to CMS standards. For example, many of CMS' policies affect hospitals, providers, and the public. It is therefore imperative that these powerful organizations leverage data science for achieving better insights, especially since much of healthcare can stand to gain improvements from new policies set forth by organizations such as these.

### Data Science and the Data Scientist

To fully tap the potentials of data science, the health care field must develop a sector of well-qualified data science specialists focused on health care data issues. The field of data science benefits from recruiting individuals that have unique data mining and analytical skills. Individuals that are interested in the field of data science should also have an in-depth understanding of data science techniques and concepts, especially in the domain of big data. Data science area concerns techniques for the extraction of information from various data, with a specific emphasis on 'Big' data displaying 'V' attributes such as veracity, value, variety, velocity, and volume (Maneth & Poulouvassilis, 2016).

Data scientists possess an in-depth understanding of data science concepts and the necessary skill sets and knowledge to utilize data science techniques. There are a number of hallmarks of an effective data science practitioner, which should inform the successful future development of the health care data science sector. First, data scientists collect data, manipulate it in a tractable form, tell the tale and present the tale to others (Loukides, 2011). In an effort to "traditionally" define the term *data scientist*, authors Liu, et al., (2009), proposed a tentative definition as a scientist committed to the study of data collection, analysis, metadata, rapid retrieval, archiving, sharing, mining to discover unexpected information and data relationships, two- and three-dimensional visualization including movement and management. Second, data scientists are normally familiar with toolkits popular in data science such as Python, Perl, R studio, Hadoop, SQL, machine learning software, and the like. Open source software, such as the R statistics kit, Python, and Perl are used by one in five data science professionals (Fayyad, 2012). Third, the data scientist benefits from artistic skills in the data science profession because it allows them to help paint a picture from the phenomena in the data (Loukides, 2011). A data scientist should have technical expertise, be curious and clever, and have the ability to tell a story through data (Patil, 2011). A data scientist should have the capacity to take an issue and incorporate multiple solutions for the different difficulties of the major problem at hand (Loukides, 2011). The skills necessary for a data scientist can vary in range. That is, a data scientist possesses skills acquired in computer science or mathematics.

Finally, in addition, a data scientist should be familiar with the four A's of data, which are architecture, acquisition, analysis, and archiving. Ultimately, it is important to note that data scientists combine creativity with persistence, the desire to incrementally create data items, the ability to experiment and the ability to iterate on a solution (Loukides, 2011). Data scientists also benefit by skills in the following areas: a) the capacity to learn the application domain, b) the ability to communicate with data users, c) attentive insight into the big picture of a complex system, d) knowledge of how data can be represented, transformed and analyzed, e) the capacity to visualize and present data, f) attention to quality, and g) ethical reasoning abilities (Stanton & De, 2013).

MODELS

Applications to healthcare must recognize that the essential components and processes of today’s data science can be found in two generally accepted models. A data science project life cycle (Data Science Central) 2014 was proposed with 7 components, as follows: 1. acquisition of data, 2. preparation of data, 3. model and hypothesis building, 4. interpret and evaluate, 5. implementation, 6. operationalize, and 7. optimize (Manna, 2014).

Another model that shows an overview of the data science process was developed by Cielen, Ali, and Meysman (2016), which propose six steps, as follows: 1. research goal setting, 2. data retrieval, 3. preparing the data, 4. exploring the data, 5. modeling the data, and 6. automating and presenting the data (Cielen, Ali, & Meysman, 2016). Table 1 Data science analysis process in Table 1 delineates the steps that build upon that foundation.

Each major step in the data science process model is comprised of goals and other processes, each respective to their major step, as shown in Table 1. Data science utilizes advanced methods to help determine predictions from the data used (Fayyad, 2012). Figure 1 shows the decisions that a data scientist undertakes when approaching data and the data scientist starts at the top of the figure making decisions that branch down to the granular level in each of the paths. As these two models are the dominant, organizing conceptual schema of the data science discipline, the development of health care data science applications must expect to map health care information needs onto their general outlines. Figure 1 developed from (Cielen, Ali, & Mysman, 2016) delineates the data science process steps map components.

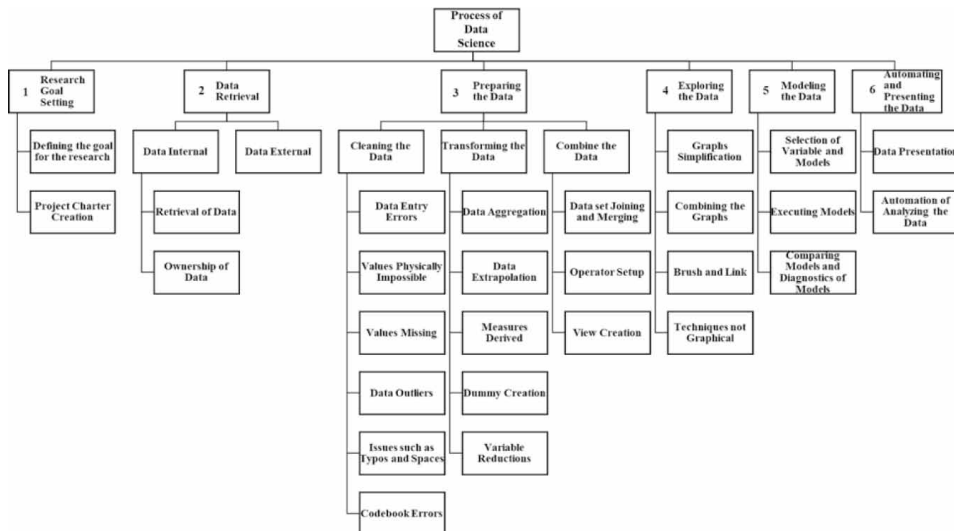
DATA SCIENCE IN HEALTH CARE

A high demand for data scientists in the field of healthcare has emerged and in the last 10 years, the information collected in healthcare systems has increased, making Big Data in healthcare possible (De la Torre Diez, Cosgaya, Garcia-Zapirain, & Lopez-Coronado, 2016). In response, healthcare needs new models to make information fully meaningful and actionable. Data scientists can contribute new knowledge to building innovative solutions that ultimately help all stakeholders in healthcare, from the patient to the treating physicians (Adam, Wieder, & Ghosh, 2017). Data science allows for the construction of data-driven theories conducive to advanced analytics in the healthcare field (Cao, 2017). One advantage of using data science processes, such as machine learning and graph analytics for deciphering big data, is that analyzing large health datasets can help in the prediction of patient outcomes. This, in turn, allows for the right clinical interventions to occur, and for new insights to surface for higher quality health care outcomes (Adam, Wieder, & Ghosh, 2017). One goal for data

Table 1. Data science analysis process

Preprocessing Steps	1. Goal Setting
	2. Obtain Data
	3. Data cleaning and formatting
Analysis Steps	4. Data exploration and summary
	5. Analytical methods
	6. Modeling
	7. Data automation and operationalization
Interpretation	8. Presentation
	9. Discussion and interpretation

Figure 1. Data Science Process Steps Map



science in healthcare is to extract new insights that will support better decisions, leading to reduced costs and the improvement of targeted quality of care for patients (Adam, Wieder, & Ghosh, 2017).

Furthermore, data science can be applied to the integrated analysis of data across fields related to health care. For example, collaboration among disciplines such as healthcare, computing, and informatics can produce innovations in data-driven theory and data-driven economy (Cao, 2017). It is essential, however, that fully trained data scientists undertake the operation of data science software in such collaborations. In this way, data scientists can help in decision-making, and leaders working in the health care industry can benefit from the insights extracted by data scientists after careful analysis of their data (Power, 2016).

Health information and health data analysis have been central to the health care sector for many years. In most cases, before the electronic health record system era, patient data were being assessed by providers, but unfortunately, the analysis was limited due to the lack of technological capacity. As is the case today, providers' goal was to improve the health of patients, but that presented challenges, such as an overload of information that could possibly be missed during initial assessment of the patient. This challenge helped to set the stage for the creation and use of electronic health record systems. Additionally, the United States Congress has been involved in marketing the use of health information technologies since 2004, when Congress began to introduce bills for the utilization of health information technologies (HIT) and electronic health information exchange systems (HIE) (Marchibroda, 2007).

Some states have made the use of such technology a top priority. This is an important step in health care, primarily because in the field of data science, most data comes from a repository or database system of some sort. The state of New York has determined there are benefits to healthcare following full adoption of HIT and HIE. In 2006, in support of the state's hope for adoption by the healthcare community, the state of New York initiated the Healthcare Efficiency and Affordability Law for New Yorkers (HEAL NY), a grant-based program that focuses on three things: 1) electronic health record (EHR) adoption, 2) electronic prescribing (ePrescribe), and the development and implantation of clinical data exchanges throughout the community (Kern & Kaushal, 2007).

HIT has allowed for the collection of protected health information (PHI). Such information includes information surrounding socioeconomic status, sexual orientation, religion, location, race,

ethnicity, gender, and mental health. Collection of such information can prepare for focused datasets that can allow for applications of data science to help determine disparities in health among types of groups in the dataset population.

## HEALTH DISPARITIES

The quality of care and outcomes in health deteriorate when there are disparities in elements such as socioeconomic status, race or ethnicity, all of which can be devastating and costly to public health. Outcomes in health are affected not only by cultural ignorance and callousness by health practitioners, but more broadly by social and economic inequities within the habitat of the population (Demeester et al., 2017). Health dissimilarities or differences that are associated with disadvantages in social, economic, and environmental settings are known as health disparities.

People are typically affected negatively in their health because of the disparate challenges they encounter around race, religion, income status, gender, age, mental health, and the like (Office of Disease Prevention and Health Promotion, n.d.). Social disadvantages are usually associated with structured differences in the healthcare system that tend to lead to health disparities (West et al., 2017). For many years people in America have tended to suffer in their health due to disparities in income, education, race, and location. Recently, there has been an effort at local, state, and regional levels to reformulate healthy standards through various determinants of health efforts (Trujillo & Plough, 2016). The Institute of Medicine has deemed such inequalities in the services and outcomes provided by health organizations as key issues to address. Contributions to such health disparate circumstances are influenced by factors in the healthcare system, such as factors in that exist in the elements of culture, provider, and those of the patient (McQuaid & Landier, 2017).

In efforts to address health disparities, health organizations have intensified their approach to social determinants of health (SDOH). SDOH is defined by the World Health Organization (WHO) as conditions in living specific to a person's environment made up of components such as birthplace, habitat or neighborhood life, age, and other factors that contribution to such conditions of living. The intensified approach by health organizations target lowering negative threats to health and focus on enhancing positive outcomes in health (Hughes et al., 2019). Healthcare is faced with excessive costs in healthcare services and such services can become wasteful, inefficient, and ineffectual due to the disparities that exist in health (King, 2016; Chin, 2016). Past studies examining disadvantaged groups have included the recognition of components that tend to influence disparities in outcomes and access in health care. Disparities in health permeate and continue in diverse type of infirmities and become expensive to health organizations.

## HEALTH EQUITY

Addressing health disparities through mitigation efforts leads to improving health equity (Anderson et al., 2018). Health equity can be defined as health excellence achieved through the eradication of disparities in health (Office of Disease Prevention and Health Promotion, n.d.). Therefore, in an effort to pursue improvements in health equity, the use of data science in healthcare should be to aid in the reduction of health disparities. The use of data science software can help analyze factors associated to health disparity. It can also aid healthcare organizations such as hospitals, clinics, provider practices, community, and public health officials find common health disparities that can help emphasize possible interventions for mitigation purposes. Although the following evaluation does not specifically treat health care data, it evaluates a number of software applications suitable to the kinds of data science operations healthcare organizations need to undertake to address issues such as health disparities.

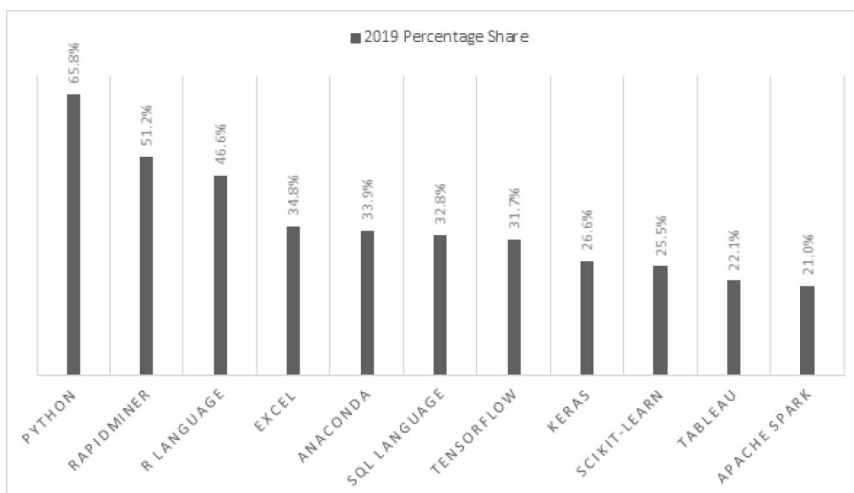
## METHODOLOGY AND DATA SOURCES

KDnuggets is a top influential site for artificial intelligence, data science, and machine learning and has received numerous academic citations (KDnuggets, 2020). An assessment of data science software was conducted in the study using KDnuggets data. Figure 2 presents how several software products reflect the available programs in data science. Among the software in the table, only software included in KDnugget's poll that categorically pertained to analytics, data science, and machine-learning software with a 30 percent or greater percentage share during the poll year 2019 were selected for the study. The poll conducted by KDnuggets sought to identify and measure utilization of analytical, data science, and machine learning software among the participants polled. The goal of the approach for this study is to use the top utilized software identified by KDnuggets to conduct an assessment of the criteria that should be present to leverage data science processes through the utilization of data science software that may be used to address health disparities. Data for Figure 2 is sourced from (Piatetsky, 2019).

Subsequently, the study incorporated a software selection criteria framework based on the following criteria elements: performance, functionality, auxiliary task support, software quality characteristics, critical vendor criteria, and software and hardware criteria (Bhargava, 2013). Sub-criteria were modified in an effort to meet the needs for data science software assessments. Although this model was originally created for data mining software, we found the framework applicable to data science software. Table 2 delineates the sub-criteria assessed for each categorical segment of the data science data selection criteria framework.

Additionally, a project management software scoring model was adopted and the scoring criteria was modified to align with evaluation needs for data science software. The original software scoring model were comprised of scores of 1 to 4 and included performance indicators of poor (1), bad (2), good (3), and excellent (4) (Gharaibeh, 2014). Tables 3 thru 6 are briefly mentioned and described in more detail prior to their insertion in the manuscript. Table 3 exhibits the new scoring model modified for data science software scoring. Table 4 exhibits the breakdown of the full assessment scored. Table 5 shows the number of functionality requirements met by functionality type. Total scores in Table 6 exhibit efforts to rank software derived from total scores.

Figure 2. Software 2019 Percentage Share



**Table 2. Software Selection Sub-Criteria**

Performance	Functionality	Auxiliary Task Support	Software Quality Characteristics	Critical Vendor Criteria	Software and Hardware Criteria
<b>Sturdiness</b>	Openness	Data Cleansing	Vertical Solution	User manual & tutorial/training	Internal and external memory
<b>Time Behavior</b>	Completeness	Data Filtering	Interface Type	Maintenance and upgrading	Source Code
	Adaptability	Binning	DBMS Standard	Consultancy	
	Interoperability	Record Deletion	Error Reporting	Product Established	
	Procedures	Handling Blanks	User Interface	Indirect Benefits	
	Security Levels		Technique Suite		
	Simultaneous users		Graphic Capabilities		
	Big Data Processing		Data Visualization		
	Data Sampling		Platform Independence		
			Platform Variety		
			Action History		
			Ease of Use		
			Domain Variety		
			Technical Source		

**Table 3. Data Science Software Scoring Model**

Score	Performance	Condition
<b>1</b>	Poor	None
<b>2</b>	Ok	Partial
<b>3</b>	Good/Excellent	Full

## RESULTS

Table 4 exhibits the individual results for the data science software evaluation based on the following new framework criteria and sub-criteria. It is important to note that Anaconda is a distributor platform and does not necessarily compute data science algorithms. However, it is an important part of a data scientist's toolkit and can facilitate and integrate other important data science applications into its platform. This limitation resulted in Anaconda's score to be lower than the other data science applications evaluated.

There are 6 total categories and a total of 38 sub-criteria categories. Table 5 shows the results for sub-criteria as it pertains to the total number of types of functionality met. Python and Tensorflow met

Table 4. Data Science Software Selection Criteria Framework

Criteria	Criteria Group	Criteria Meaning	Python	RapidMiner	R Language	Excel	Anaconda	SQL Language	Tensorflow
Performance									
Sturdiness	Reliability	Performs without crashes	3	3	3	3	3	3	3
Time Behavior	Efficiency	Speed of computational results	3	2	3	2	2	3	3
Functionality									
Openness	Functional	Accessible for more development	3	3	3	1	2	3	3
Completeness	Functional	The extent of software required functions met	3	3	3	2	2	3	3
Adaptability	Functional	Customizable for industries or companies	3	3	3	2	2	3	3
Interoperability	Functional	Capacity to integrate with other applications	3	2	3	2	3	3	3
Procedures		Has suite of procedures for data science	3	3	3	3	2	2	3
Security Levels	Functional	Policy exists for security application of software such as identification of users and encrypting data	3	3	3	3	2	3	3
Simultaneous users	Functional	Can handle simultaneous users on the system	2	2	2	3	3	1	2
Data type Flexibility		Supports a variation of types of data	3	3	3	2	2	3	3
Big Data Processing		Capacity for processing high data volumes	3	3	3	1	2	3	3
Data sampling		Data sampling capacity at random for predictive models	3	3	3	3	2	3	3
Auxiliary Task Support									
Data Cleansing		Data modification of values for cleaning data	3	3	3	3	2	3	3
Data Filtering		Capacity to filter data based on a set of selections defined by user	3	3	3	3	2	3	3
Binning		Improved efficiency by allowing binning of data that is continuous	3	3	3	3	2	3	3
Record Deletion		Biased or unbiased record deletion capacity	3	3	3	3	2	3	3
Handling blanks		Blank handling capacity on entries	3	3	3	3	2	3	3
Software Quality Characteristics									
Vertical Solution	Personalization	Software package customized version to help meet specific industry requirements	3	3	3	3	3	3	3
Interface type	Personalization	Package type is user interface based	3	3	3	3	3	3	3
DBMS standard	Portability	Other types DB software packages such as SQL server and Oracle can be accessed by the software	3	3	3	3	3	3	3
Error reporting	Usability	Ability to message and report on errors	3	3	3	3	3	3	3
User interface	Usability	User interface ease of utilization	2	3	2	3	3	2	2
Technique Suite		Capacity to employ techniques such as time series and modeling	3	3	3	3	3	2	3
Graphic Capabilities		High graphic visualization quality for viewing such as decision trees	3	3	3	3	2	2	3
Data visualization	Usability	Effective data representation capacity	3	3	3	3	2	2	3
Platform Independence		Capacity to add other models and/or functionalities	3	2	3	3	3	2	3
Platform variety	Portability	Software can be used on a variety of platforms	3	2	3	3	2	3	3
Action history	Usability	In data science processes, software allows to modify action history	3	3	3	3	2	3	3
Ease of use	Usability	Users can easily learn and operate the software	3	2	2	3	3	2	2
Domain variety	Usability	Software is domain diverse and capable of being tailored to other industry for business problem solving	3	3	3	3	3	3	3
Technical Source	Opinion	Other vendors and in-house experts and consultants opinion on software	2	2	2	2	3	2	2
Critical Vendor Criteria									
User manual & tutorial/training	Vendor	Manuals, guidelines, tutorials, and other learning material available to users	3	3	3	3	3	3	3
Maintenance and upgrading	Vendor	Contracts and available for upgrades based on annual agreement as maintenance program	3	3	3	3	3	3	3
Consultancy	Vendor	Technical support availability to users	2	2	2	3	2	2	2
Product Established		Maturity of the software product	2	2	2	3	2	3	2
Indirect benefits	Benefits	Customer service improvement	1	2	1	2	1	1	1
Software and Hardware Criteria									
Internal and external memory	Hardware	Package run based on storage that is primary and secondary	2	2	2	2	3	2	3
Source code	Software	Source code availability	3	3	3	1	3	3	3

Table 5. Number of Functionality Requirements Met

Type Functionality	Python	RapidMiner	R Language	Excel	Anaconda	SQL language	Tensorflow
Full	31	27	30	27	17	26	31
Partial	6	11	7	8	20	10	6
None	1	0	1	3	1	2	1

the highest number of full functionality sub-criteria components, followed by R language, Rapidminer and Excel, SQL language, and Anaconda. Among partial functionality types, Anaconda had the highest met followed by Rapidminer, SQL language, Excel, R language, Python and Tensorflow. For those with no functionality, the highest number met was Excel followed by SQL language, a tie among Python, R language, Anaconda, and Tensorflow. Rapidminer had zero in this category.

Table 6 exhibits the overall scored results for each software ranked from highest to lowest.

The highest rank software programs exhibited in Table 6 indicate that Tensorflow and Python met the majority of the sub-criteria components. There were no major differences between tensorflow and python. R language was ranked second followed by RapidMinder, SQL language and Excel, and Anaconda. There were no meaningful differences noted between the top four software ranked software based on their capacity to analyze structured and unstructured data. Although a powerful data extractor and data manipulator language, the SQL language in comparison to the four top-ranked software, did not fully meet the technique suite sub-criteria and lacked in data visualization capabilities. However, SQL should be integrated with software platforms to optimize data processes important to data science workflows. Excel showed to be a competitor among the software assessed but lacked in its capacity to fully allow big data processing and it is not considered an open source software limiting valuable contributions from the development community. As noted earlier, Anaconda is a distribution platform and acts as a gateway platform to multiple data science software. Although it scored the lowest due to only meeting partial functionality criteria through its capacity to integrate software to its platform, it is worth noting that it allows for better efficiencies and access to data science software.

## CONCLUSION

The field of data science utilizes various methodological approaches for analyzing data in any domain or sector, including healthcare. The healthcare sector has not seen the full benefits of data science. However, this sector is beginning to dive into the field to explore new algorithms and methods that will aid in higher quality of care and quality outcomes. With the creation of new technologies and their capacities of creating data, possibilities into predicting probable outcomes based on historical data are now possible (Spruit & Lytras, 2018). Such innovations are especially likely, as this paper has argued above, in relation to healthcare sector networks connected through CMS and state initiatives such as HEAL NY.

The evaluation insights gained from this study based on the Data Science Software Selection Criteria Framework delineate how data science functionalities can help aid healthcare in approaching analytical processes with new analytical applications suitable for healthcare. For example, based on

**Table 6. Top Ranked by Score Total**

<b>Software</b>	<b>Total Score</b>
<b>Tensorflow</b>	<b>106</b>
<b>Python</b>	<b>106</b>
<b>R Language</b>	<b>105</b>
<b>RapidMiner</b>	<b>103</b>
<b>SQL Language</b>	<b>100</b>
<b>Excel</b>	<b>100</b>
<b>Anaconda</b>	<b>92</b>



the highest ranked software in the study, Tensorflow and Python both have the capacity of automating and modeling the analysis of variables such as income, education, race, age, and cross-referencing such variables to outcomes in patient care and finance to determine outcomes that reveal health disparities. This paper documents a process that provides an opportunity to address health disparities. Rankings should constantly be revisited due to advancements and development of new software and changes within the discipline of data science. Furthermore, contributions in this work allow the healthcare community to continually and iteratively evaluate data science software, as progressions are made, using the methods in this research.

This paper has demonstrated the data science capabilities through exhibiting the potential utility of leading software to perform the kinds of data science operations that can achieve improved care within such networks by addressing such factors as health disparities.

## REFERENCES

- Adam, N. R., Wieder, R., & Ghosh, D. (2017). Data science, learning, and applications to biomedical and health sciences. *Annals of the New York Academy of Sciences*, 1387(1), 5–11. doi:10.1111/nyas.13309 PMID:28122121
- Allen, C., Tsou, M., Aslam, A., Nagel, A., & Gawron, J. (2016). Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. *PLoS One*, 11(7), e0157734. doi:10.1371/journal.pone.0157734 PMID:27455108
- Anderson, A. C., O'Rourke, E., Chin, M. H., Ponce, N. A., Bernheim, S. M., & Burstin, H. (2018). Promoting health equity and eliminating disparities through performance measurement and payment. *Health Affairs*, 37(3), 371–377. doi:10.1377/hlthaff.2017.1301 PMID:29505363
- Baptista, M., Vasconcelos, J. B., Rocha, Á., Silva, R., Carvalho, J. V., Jardim, H. G., & Quintal, A. (2019). The impact of perioperative data science in hospital knowledge management. *Journal of Medical Systems*, 43(2), 41. Advance online publication. doi:10.1007/s10916-019-1162-3 PMID:30637593
- Bhargava, N., Aziz, A., & Rajiv, A. (2013). Selection criteria for data mining software: A study. *IJCSI International Journal of Computer Sciences*, 10(3).
- Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys*, 50(3), 1–42. doi:10.1145/3076253
- Chase, J. D., & Vega, A. (2016). Examining health disparities using data science. *Research in Gerontological Nursing*, 9(3), 106–107. doi:10.3928/19404921-20160404-01 PMID:27210530
- Chin, M. H. (2016). Creating the business case for achieving health equity. *Journal of General Internal Medicine*, 31(7), 792–796. doi:10.1007/s11606-016-3604-7 PMID:26883523
- Cielen, D., Ali, M., & Meysman, A. (2016). *Introducing data science: Big data, machine learning, and more, using Python tools*. Manning.
- De la Torre Díez, I., Cosgaya, H. M., Garcia-Zapirain, B., & López-Coronado, M. (2016). Big data in health: A literature review from the year 2005. *Journal of Medical Systems*, 40(9), 209. Advance online publication. doi:10.1007/s10916-016-0565-7 PMID:27520614
- Delaney, C. W., & Westra, B. (2016). Big data. *Western Journal of Nursing Research*, 39(1), 3–4. doi:10.1177/0193945916671687 PMID:30208772
- DeMeester, R. H., Xu, L. J., Nocon, R. S., Cook, S. C., Ducas, A. M., & Chin, M. H. (2017). Solving disparities through payment and delivery system reform: A program to achieve health equity. *Health Affairs*, 36(6), 1133–1139. doi:10.1377/hlthaff.2016.0979 PMID:28583973
- Detmer, D. E., & Shortliffe, E. H. (2014). Clinical informatics. *Journal of the American Medical Association*, 311(20), 2067. doi:10.1001/jama.2014.3514 PMID:24823876
- Dunn, M. C., & Bourne, P. E. (2017). Building the biomedical data science workforce. *PLoS Biology*, 15(7), e2003082. doi:10.1371/journal.pbio.2003082 PMID:28715407
- Erdman, S. L. (2020, May 6). *Black communities account for disproportionate number of COVID-19 deaths in the US, study finds*. CNN. <https://www.cnn.com/2020/05/05/health/coronavirus-african-americans-study/index.html>
- Fayyad, U. (2012, July 4). *Data science revealed: A data-driven glimpse into the burgeoning new field*. <https://fayyad.com/data-science-revealed-a-data-driven-glimpse-into-the-burgeoning-new-field/>
- Gharaibeh, H. M. (2014). Developing a scoring model to evaluate project management software packages based on ISO/IEC software evaluation criterion. *Journal of Software Engineering and Applications*, 07(01), 27–41. doi:10.4236/jsea.2014.71004
- Hughes, M. C., Baker, T. A., Kim, H., & Valdes, E. G. (2019). Health behaviors and related disparities of insured adults with a health care provider in the United States, 2015–2016. *Preventive Medicine*, 120, 42–49. doi:10.1016/j.ypmed.2019.01.004 PMID:30639668
- KDnuggets. (n.d.). *About KDnuggets*. <https://www.kdnuggets.com/about>

- Kern, L. M., & Kaushal, R. (2007). Health information technology and health information exchange in New York State: New initiatives in implementation and evaluation. *Journal of Biomedical Informatics*, 40(6), S17–S20. Advance online publication. doi:10.1016/j.jbi.2007.08.010 PMID:17945542
- King, C. (2016). Disparities in access to preventive health care services among insured children in a cross sectional study. *Medicine*, 95(28), e4262. doi:10.1097/MD.00000000000004262 PMID:27428239
- Liu, L., Zhang, H., Li, J., Wang, R., Yu, L., Yu, J., & Li, P. (2009). Building a community of data scientists: An explorative analysis. *Data Science Journal*, 8, 201–208. doi:10.2481/dsj.008-004
- Loukides, M. (2011) *What is data science?* O'Reilly Media. <https://www.oreilly.com/data/free/what-is-data-science.csp>
- Maneth, S., & Poulouvassilis, A. (2016). Data science. *The Computer Journal*, 60(3), 285–286. doi:10.1093/comjnl/bxw073
- Manna, M. (2014, December 18). *The data science project lifestyle*. Data Science Central. <https://www.datasciencecentral.com/profiles/blogs/the-data-science-project-lifecycle>
- Marchibroda, J. M. (2007). Health information exchange policy and evaluation. *Journal of Biomedical Informatics*, 40(6), S11–S16. Advance online publication. doi:10.1016/j.jbi.2007.08.008 PMID:17981099
- McQuaid, E. L., & Landier, W. (2017). Cultural issues in medication adherence: Disparities and directions. *Journal of General Internal Medicine*, 33(2), 200–206. doi:10.1007/s11606-017-4199-3 PMID:29204971
- Murphy, W. F., Murphy, S. S., Buettner, R. R., & Gill, G. (2015). Case study of a complex informing system: Joint interagency field experimentation (JIFX). *Informing Science: The International Journal of an Emerging Transdiscipline*, 18, 63–109. 10.1093/comjnl/bxw07310.28945/2289
- O'Connor, S. (2018). Big data and data science in health care: What nurses and midwives need to know. *Journal of Clinical Nursing*, 27(15-16), 2921–2922. doi:10.1111/jocn.14164 PMID:29148112
- Office of Disease Prevention and Health Promotion. (n.d.). *Disparities*. HealthyPeople.gov. <https://www.healthypeople.gov/2020/about/foundation-health-measures/Disparities>
- Patil, D. J. (2011). *Building data science teams*. O'Reilly Media.
- Piatetsky, G. (2019). *Python leads the 11 top data science, machine learning platforms: Trends and analysis*. KD Nuggets. <https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html>
- Power, D. J. (2016). Data science: Supporting decision-making. *Journal of Decision Systems*, 25(4), 345–356. doi:10.1080/12460125.2016.1171610
- Press, G. (2013, May 28). *A very short history of data science*. Forbes. <https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#1161694a55cf>
- Rumbold, J. M., & Pierscionek, B. K. (2017). A critique of the regulation of data science in healthcare research in the European Union. *BMC Medical Ethics*, 18(1), 27. Advance online publication. doi:10.1186/s12910-017-0184-y PMID:28388916
- Smith, F. J. (2006). Data science as an academic discipline. *Data Science Journal*, 5, 163–164. doi:10.2481/dsj.5.163
- Spruit, M., & Lytras, M. (2018). Applied data science in patient-centric healthcare: Adaptive analytic systems for empowering physicians and patients. *Telematics and Informatics*, 35(4), 643–653. doi:10.1016/j.tele.2018.04.002
- Stanton, J. M., & De, G. R. (2013). *An introduction to data science*. Sage (Atlanta, Ga.).
- Trujillo, M. D., & Plough, A. (2016). Building a culture of health: A new framework and measures for health and health care in America. *Social Science & Medicine*, 165, 206–213. doi:10.1016/j.socscimed.2016.06.043 PMID:27405727
- West, K. M., Blacksher, E., & Burke, W. (2017). Genomics, health disparities, and missed opportunities for the nation's research agenda. *Journal of the American Medical Association*, 317(18), 1831. doi:10.1001/jama.2017.3096 PMID:28346599

Westra, B. L., Sylvia, M., Weinfurter, E. F., Pruinelli, L., Park, J. I., Dodd, D., Keenan, G., Senk, P., Richesson, R., Baukner, V., Cruz, C., Gao, G., Whittenburg, L., & Delaney, C. W. (2017). Big data science: A literature review of nursing research exemplars. *Nursing Outlook*, 65(5), 549–561. doi:10.1016/j.outlook.2016.11.021 PMID:28057335

Zhang, X., Pérez-Stable, E. J., Bourne, P. E., Peprah, E., Duru, O. K., Breen, N., Berrigan, D., Wood, F., Jackson, J. S., Wong, D. W. S., & Denny, J. (2017). Big data Science: Opportunities and challenges to address minority health and health disparities in the 21st century. *Ethnicity & Disease*, 27(2), 95. doi:10.18865/ed.27.2.95 PMID:28439179

*Jose O. Huerta is an experienced healthcare information technology professional with over 20 years of experience in the healthcare information technology (HIT) space. He holds a Master of Science (MS) degree in business management from Troy University and a Bachelor of Science (BS) degree in business management from Park University. He holds two certifications in electronic health record technology from the American Health Information Management Association (AHIMA). He is a board member of the Texas Health Information Management Association (TXHIMA) and received his PhD in May 2021 at the University of North Texas in Denton, Texas.*

*Gayle Prybutok is an Assistant Professor, Health Services Administration in the Department of Rehabilitation and Health Services, College of Health and Public Service, at the University of North Texas. She served as the Coordinator for the Health Services Administration Master's Degree and was instrumental in creating that program and the Ph.D. in Health Services Research. Dr. Gayle Prybutok holds a Bachelor's in Nursing from Thomas Jefferson University, an MBA from Texas Woman's University and a Ph.D. in Information Science with a focus on Health Informatics from the University of North Texas. She formerly served as the Chief Nursing Officer of a local hospital, Director of home health and hospice agencies, and was the Executive Director of a national non-profit funded by NIH to procure human tissue for research. Her research interests include online health communication, health care quality improvement, and health education via the Internet.*

*Victor R. Prybutok is a Regents Professor of Decision Sciences in the Information Technology and Decision Sciences Department in UNT's G. Brint Ryan College of Business, Vice Provost for Graduate Education, and Dean of the Toulouse Graduate School at the University of North Texas. He received, from Drexel University, his B.S. with High Honors in 1974, an M.S. in Bio-Mathematics in 1976, an M.S. in Environmental Health in 1980, and a Ph.D. in Environmental Analysis and Applied Statistics in 1984. Dr. Prybutok is an American Society for Quality certified quality engineer, certified quality auditor, certified manager of quality / organizational excellence, and an accredited professional statistician (PSTAT®) by the American Statistical Association. He has authored over 200 journal articles, more than 350 conference presentations/proceedings, and several book chapters. In 2017 he received the American Society for Quality Gryna Award for a co-authored manuscript and in 2018 was awarded the Decision Sciences Institute Lifetime Distinguished Educator Award. Most recently he was awarded the 2020 Distinguished Service Award by the Southwest Region of the Decision Sciences Institute.*

# Big Data Applications in Vaccinology

Joseph E. Kasten, Pennsylvania State University, York, USA

## ABSTRACT

The development of vaccines has been one of the most important medical and pharmacological breakthroughs in the history of the world. Besides saving untold lives, they have enabled the human race to live and thrive in conditions thought far too dangerous only a few centuries ago. In recent times, the development of the COVID-19 vaccine has captured the world's attention as the primary tool to defeat the current pandemic. The tools used to develop these vaccines have changed dramatically over time, with the use of big data technologies becoming standard in many instances. This study performs a structured literature review centered on the development, distribution, and evaluation of vaccines and the role played by big data tools such as data analytics, data mining, and machine learning. Through this review, the paper identifies where these technologies have made important contributions and in what areas further research is likely to be useful.

## KEYWORDS

Big Data, Machine Learning, Pharmaceutical, Structured Literature Review, Vaccine, Vaccinology

## INTRODUCTION

As this manuscript is being prepared, the world is struggling to produce a vaccine to prevent against the Covid-19 virus. Research on a wide variety of biological topics is currently underway and these build on countless other studies that have been completed over the preceding years. Many of these research efforts have relied on advanced analytical and data manipulation tools to both identify promising vaccine targets and to evaluate the efficacy of those already administered. The purpose of the present effort is to provide a structured review of the research literature describing the use of Big Data technologies in the development and evaluation of vaccines.

These analytical tools have been used to extract actionable information from vast databases of vaccine trial results (Ackerman, Barouch & Alter, 2017), develop an understanding of the unpredicted side effects of already developed vaccines (Iqbal et al, 2015), and predict the immunological attack points for future vaccine development (Conti & Karplus, 2019). In each of these scenarios, the use of Big Data technologies has helped to create vaccines more quickly, with more efficacy, and to identify additional uses for existing vaccines, thus increasing the pool of available vaccines without the expensive and time consuming process of developing them from the ground up.

To be clear, this is not a paper devoted to reviewing the technologies specifically utilized in the quest for a Covid-19 vaccine, though many of the tools described in this project will have undoubtedly been used in this effort. Rather, it is meant to describe the many Big Data-based tools that have been used in the recent past that have brought the field of vaccinology and immunology to the point where

DOI: 10.4018/IJBDAH.20210701.oa5

This article, published as an Open Access article on April 23, 2021 in the gold Open Access journal, International Journal of Big Data and Analytics in Healthcare (converted to gold Open Access January 1, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

the development of a complex and critical vaccine such as those coming out for the Covid-19 effort can be developed, tested, and administered within the space of a single year. As in so many medical breakthroughs, what looks like a standalone discovery is actually the result of years of methodical, basic research that has paved the way for a specific advancement to happen in the manner that it has. Thus, this review of the literature describes the efforts that have led up to the present successes. The contribution made by a paper of this type is two-fold. First, it provides a framework within which researchers in the field, or those considering entering it, can place their research. This paper provides, for these researchers, a good overview of what is ongoing and, even more importantly, what still needs to be done. Secondly, for practitioners this paper represents a guide to those technologies that are being used or under development and might provide them with ideas about how to improve their processes or even develop new approaches to using big data in vaccinology.

## BACKGROUND

When the non-epidemiologist thinks of vaccines, they think of flu shots and other immunizations given routinely to children and young adults. But, underlying these medical achievements is the need for processing and drawing understanding from huge datasets. Many of these Big Data processes are similar to those used in other industries, so from that perspective vaccine development and evaluation draws on tools that have matured in other areas such as finance and supply chain (Jordan, Dossou & Chang Jr., 2019). It is important, though, to understand how these tools are being employed, for as this understanding is more broadly dispersed throughout the data community, new applications can be more readily devised and more opportunities for protection from these illnesses and parasites can be developed.

Before proceeding with the study, it is important to define what is meant, in this effort, by Big Data technologies. This is important because of the wide range of understanding that exists about the definitions of Big Data and the tools and technologies associated with it. For the present project, a rather wide view of these terms is taken. This will allow the results to paint as clear a picture as possible of the state of the art and reduce the number of studies overlooked for terminological reasons. For the purposes of this study, the term “Big Data technologies” will include all forms of data analytics, machine learning, deep learning, and datamining. To reduce the complexity of the searches, the study does not use all possible terms describing these processes. However, the databases searched do a very good job of providing results that include other terms related to the basic terms listed above, so for example papers including terms such as “artificial intelligence” and “support vector machine” will be included even without those terms being used in the actual search expression. The terms used in this study are defined below:

- Data mining: using software and statistical models to search for patterns in large datasets.
- Machine Learning (ML): A broad category of algorithms that use data to train a model to make decisions. This includes a very wide collection of models and tools that are beyond the scope of this paper but can be reviewed in (Obermeyer & Emmanuel, 2016).
- Deep Learning: A branch of machine learning that enables tools such as Artificial Neural Networks (ANN). These can be trained to make decisions that support such processes as classification or speech recognition. These are closely related to big data technologies because they require very large datasets to train and verify the algorithms.
- Data Analytics: A broad term covering all of the statistical, visual, and mathematical tools and techniques commonly used to analyze and draw meaningful information from big data repositories.

While this is not an exhaustive list of technologies related to Big Data, it provides a good starting point with which to examine the literature surrounding the use of Big Data technologies in

the development and evaluation of vaccines. The following section details the methodology used in this structured literature review.

## METHODOLOGY

As this is a structured review of the literature surrounding the use of Big Data technologies in the vaccine development and evaluation space, it is crucial to employ a methodology that ensures rigor and provides results that can be replicated. The present study follows the process outlined by Briner & Denyer (2012). These authors provide a five-step process to follow when completing a study of this sort:

- Identify the research question(s).
- Locate and select relevant studies
- Critically appraise the studies
- Analyze and synthesize the findings
- Disseminate the findings

The study seeks answers to the following research questions:

- For which vaccine development/evaluative processes are big data technologies (as defined in this study) being used to improve or analyze?
- For how long have big data technologies been used to improve or analyze vaccine development/evaluative processes?

The first research question captures the overarching reason for the study. The second research question, once resolved, will provide the reader with an understanding of the maturity level of the Big Data/vaccine combination. For researchers, this provides some insight into the availability of new research directions, while practitioners in the data science, pharmaceutical, and regulatory fields will be able to evaluate how deeply embedded these tools are, and that might lead to a better understanding of how they could fit into the overall vaccine development process.

In order to facilitate the structured review, certain “ground rules” must be in place to both ensure repeatability and to limit the scope of the review. Therefore, the following stipulations apply to this study:

- The literature included in the study will consist only of published works such as journal articles, conference proceedings, book chapters, etc. White papers, works-in-progress, and other so-called “gray” literature, while quite possibly making important contributions to the field are excluded from the study. These might be an interesting area of inquiry for future study.
- Literature generally aimed at the practitioner, rather than researchers, is also excluded from the study. As above, there is likely useful information in these publications, but the point of the current effort is to locate the areas currently being researched and developed in the academic literature. Again, this would be a fruitful area for future research.

Searches were carried out in many of the major electronic databases that provide coverage for the vast majority of scholarly journals in the fields of data analytics, information technology, and healthcare. The databases analyzed were ABI/Informs, Emerald, IEEE Explore, JSTOR, Science Direct, Scopus, Springer, Taylor & Francis, Web of Science, Google Scholar, PubMed, CINAHL, and ACM. The searches were conducted by searching the metadata (title, abstract, keywords) using the following search terms in all combinations: big data, data analytics, deep learning, machine learning,

data mining (datamining), vaccine, and immunology. These search terms cover a great deal of the topic area, but this study does not claim to be exhaustive. However, as the searching continued and the amount of duplication among the various databases increased, it seems that this search protocol identified a large percentage of the available literature. Figure 1 displays the study protocol.

The returns from the initial set of searches numbered in the thousands. The next step in the process was to identify which of the articles in the search results fit the needs of the study. In some cases, the article title provided sufficient evidence that the article fit the study. However, for many articles the title was sufficiently vague that the abstract, the introduction, and in some cases the findings and conclusion, were analyzed to finally determine that the article was indeed about the use of some form of Big Data technology in the field of vaccinology. After culling the studies that do not focus their attention on the use of Big Data technologies on the development and application of vaccines, there remained 325 papers. The composition of these papers and the research themes contained therein are described in the Findings section.

FINDINGS

There are two subsections in the Findings section. The first subsection provides descriptive statistics for the dataset collected using the protocols described above. The second subsection provides an analysis of the various themes and subthemes that are evident in the academic literature.

Descriptive Statistics

Of the 325 documents included in this study, there are only three types of documents represented: journal articles, conference proceedings, and edited book chapters. No theses, dissertations, or other types of documents were uncovered. Table 1 displays the breakdown of documents by document type.

The proportion of journal articles to conference proceedings and book chapters varied by theme but in each case the number of journal articles far exceeds the combination of proceedings and book chapters. In only two cases, the group of documents focused on pharmacovigilance and reverse vaccine design, were there no conference proceedings or book chapters. These trends will be discussed further in the Discussion section.

There was no temporal restriction put on the search terms in order to gain an understanding of when the association between Big Data technologies and vaccinology first began and how quickly it has grown. Figure 2 provides a graphical display of the growth of Big Data technologies and vaccine-related literature. A few early papers, the first being published in 1998, made some initial connections between the two concepts, but it wasn't until approximately 2008 that the body of literature began

Figure 1. Structured literature review protocol

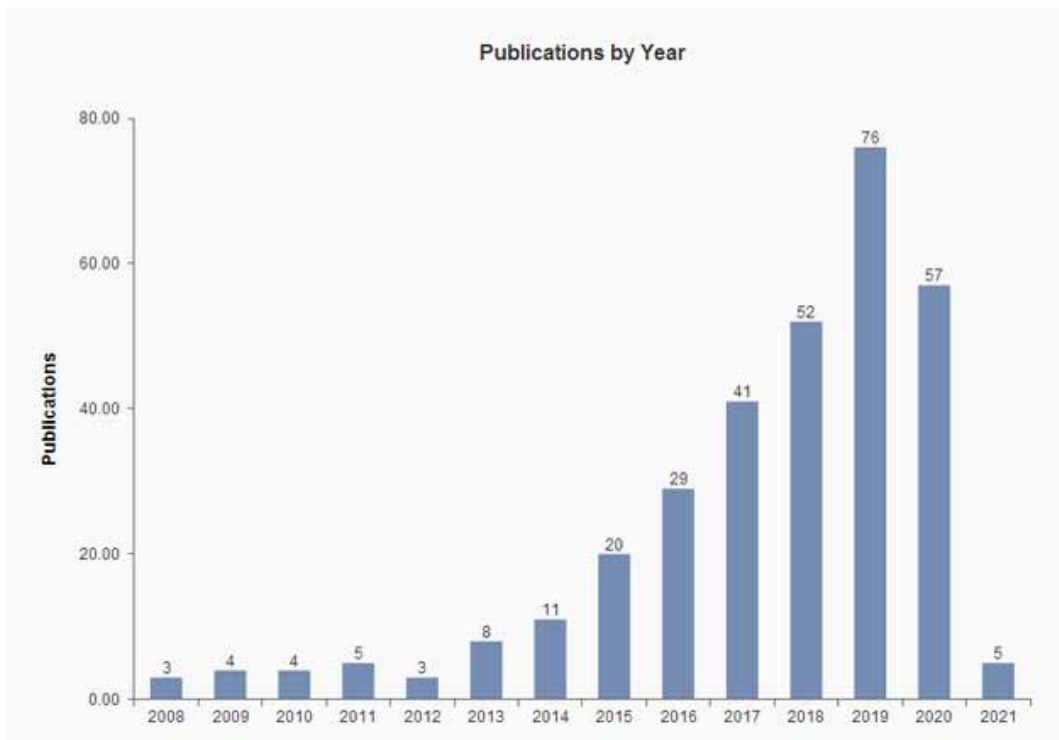


Table 1. Documents found by type

Document Type	Number Found
Journal articles	282
Conference Proceedings	34
Book Chapters	9



Figure 2. Number of publications by year



to build in earnest. A sharp increase in publications begins in 2013 and peaks in 2019. The searches were all complete by the end of November, 2020, so it is likely that the final values for 2020 will be somewhat higher. Also, a few publications scheduled for 2021 are included in the data, giving a preview of the upcoming activity.

## Research Themes

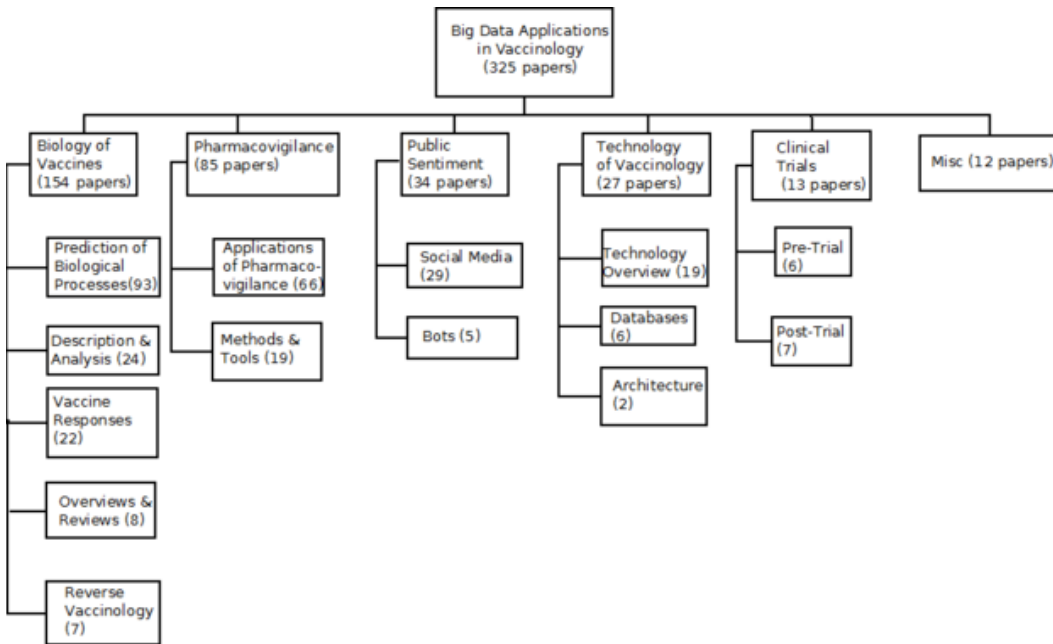
During the analysis of this collection of literature, a number of distinct research themes emerged. However, it also became apparent that even within these rather well-defined themes, there was a need to provide additional clarity regarding their individual emphases. To fulfill that need, many of the themes described in this paper also contain a number of sub-themes, or subdivisions, within them. These will provide a greater degree of specificity in providing an explanation of the six research themes identified. A graphical depiction of these research themes is presented in Figure 3.

As the scale of the current study is rather large at 325 items, it is impractical to describe each of these papers in any detail. However, it is important that some level of detailed explanation is provided to convey the importance of this body of literature. As a compromise, a selection of works from each area of concentration (sub-themes) is highlighted and discussed and a larger selection is referenced for each sub-theme in the tables that follow each sub-section.

### *Theme: Biology of Vaccines*

The first and largest theme in the body of literature centers on the actual biology of vaccines and the antigens they protect against (152 papers). Within this theme can be found a number of sub-

**Figure 3. Research themes and sub-themes**



themes, the largest of which deals with the prediction of various biological interactions among the various components of antigens and antibodies. Within this sub-theme, the largest thread centers on the prediction of how certain biological entities will interact. Epitopes, sometimes known as antigenic determinants, are the specific place on the antigen that an antibody can attach to in an effort to neutralize the attacker (Khanna & Rana, 2017). Kar et al (2018) describe various methods to identify likely immunogenic amino acid sequences such as artificial neural networks (ANN) and support vector machines (SVM). Rosyda, Adji & Setiawan (2016) use a cascading neural network to predict the epitope region on the P24 protein on the HIV virus. A deep neural network tool is used by Sher, Zhi & Zhang (2017) on a number of standard protein datasets and was shown to achieve substantially better performance than other accepted prediction tools. Additionally, Solihah, Azhari & Musdholifah (2020) couple an oversampling technique with a SVM to develop a conformational epitope prediction model and a method called BPairwise is developed by Zhang & Niu (2010) to predict variable-length epitopes.

Epitopes are composed of proteins, and the next sub-theme focuses on the identification and prediction of virulent proteins. Munteanu et al (2014) create a new model using linear discriminant analysis to discriminate the 3D structure of lectins (carbohydrate-binding proteins) from a collection of protein structures with unknown functions. These are important in developing therapies and vaccines for a wide range of cancers, parasitic infections, and other diseases. An ensemble of SVM are used by Nanni & Lumini (2009) to predict the functions of virulent mechanisms in pathogens by extracting the features directly from the protein's DNA. Singh, Singh, and Sisodia (2019) approach the limiting factor in the utilization of ML in protein prediction, the lack of adequate training datasets, by developing a composite model for a multitask learning framework. This will facilitate the building and verification of training datasets, thus opening additional pathways for protein identification and cataloging. Lastly, El-Manzalawy and colleagues (2016) tackle the problem of predicting the existence and character of surface proteins on the malaria parasite using a semisupervised ML algorithm.

Peptides are composed of a small number of amino acids chained together and are distinguished from proteins by consisting of a smaller number of amino acids, though the cutoff between the two is somewhat arbitrary. Peptides bind to the major histocompatibility complex (MHC), which are cell surface proteins that bind peptides to the cell surface so they may be recognized by T-cells (Janeway et al, 2001). They are important in the identification of cells that should be considered part of the organism or possible antigens that need to be attacked. Vang and Xie (2017) use ANN technology to predict the binding tendencies of these peptides. Boehm et al (2019) use a similar approach, but use public databases describing MHC immunopeptides to train random forest classifiers, a specialized form of ML. Liu et al (2020) use a similar approach to identify these MHC binding peptides in the course of vaccine development for SARS-CoV-2.

Predicting antigen/antibody interaction is the next area of interest in this subtheme. This is the next step after the surveys of epitope and peptide behaviors described above. These studies are also applicable to vaccine development beyond those of infectious diseases, such as those for certain cancers. For example, Smith et al (2019) use ML techniques to analyze the behavior of tumor antigen immunogenicity. Conti and Karplus (2019) use ML to uncover a universal HIV vaccine that would neutralize multiple variants of the disease. Focusing on specific antibodies, IgA and IgG, Khanna and Rana (2017) use an ensemble ML approach to predict the existence and amount of these in a sample dataset. Successful prediction of these antibodies could help formulate treatments and vaccines.

The final thread within the prediction sub-theme contains studies that seek to predict the performance, or sometimes lack of performance, of vaccines once they are administered. Examples of these include Parvandehe et al (2019) study using ML to predict the antibody response to the influenza vaccine, a study by Lee et al (2015) that seeks to understand exactly how the H1N1 influenza A virus provides protection for its human recipients, and the effort by Hemedan et al (2020), which attempts to predict the existence of vaccine-derived poliovirus outbreaks.

The second sub-theme in the biology theme is the use of Big Data technologies in the description and analysis of virus types and components. An important example of this is the use of a novel data-mining algorithm to identify the sexual development cycle of the malarial parasite in order to facilitate improved vaccine design. Mining of huge libraries of Kröhnke pyridines to find keys to the production of an arenavirus vaccine is another important application of these tools. Pyridines are organic compounds that are important in the creation of pharmaceutical compounds. Other efforts in this sub-theme involve the use of datamining samples taken from HIV-positive people to understand the exact immunological responses of their individual immune systems to see how they tried to fight off the virus, even if unsuccessfully, and a similar effort to mine the compounds from African flora with potential to inhibit the activity of the Ebola virus.

The analysis of the actual response to a vaccine, rather than a prediction, is the content of the third sub-theme. In some cases, the actual mechanism by which vaccines provide protection is not completely understood, so post-vaccination analysis can be useful in subsequent vaccine development as well as defining adjuvant therapies. Flanagan et al (2013) derive their understanding of a vaccine response by profiling the reaction of the entire immune system rather than only the cellular level of response. Pittala, Morrison, and Ackerman (2019) perform this process on humoral immune responses to HIV. Chaudhury et al (2020) analyze the effects of adjuvant formulation on vaccine-induced immunity.

Sub-theme four includes those authors whose goal is to provide an overview of the field, either as an introductory work or to provide a footing for future research. Some examples of these works include Cotugno et al (2015) effort to review techniques for gene expression analysis as part of the search for a functional cure for pediatric HIV, Izak, Klim and Kaczanowski's (2018) review of the use of bioinformatics in the definition of host-parasite interactions in the fight against malaria, and Olafsdottir, Lindqvist, and Harandi's (2015) work to define the tools used in the derivation of molecular signatures of vaccine adjuvants. These papers and others like them provide a suitable underpinning of the use of analytic tools as applied to very specific problems, just as the current project provides a larger portrait of the field for a wider view of the use of these technologies in many aspects of vaccinology.

The last sub-theme in this section is labeled “reverse vaccinology,” which is a term that has gained some traction over the past few years. In reality, much of the material reviewed in the current theme falls under the heading of reverse vaccinology (RV), which simply means that the search for epitopes and their compatible antigens occurs *in silico* (on computers) rather than *in vivo* (in living organisms). The reason for setting these works aside into a separate sub-theme is because their language represents a specific term within the field of vaccinology. This terminology might be used to a) define a specific project or effort or b) be used as a specific search term by subsequent researchers. Therefore, the topics covered in this sub-theme include the evaluation of different classification algorithms to predict protein sequences (Heinson et al, 2019), a comparison of open source RV tools for bacterial vaccine antigen discovery (Dalsass et al, 2019), and a description of how ML and RV were used in the COVID-19 vaccine development process (Ong et al, 2020). Table 2 presents the works in the Biology of Vaccines theme.

### **Theme: Pharmacovigilance**

Pharmacovigilance is the term given to those activities which allow the government and the vaccine manufacturer to monitor the performance and side effects of a vaccine, or any drug, after it is approved and released. In many cases, the monitoring is done with data submitted to the Vaccine Adverse Event Reporting System (VAERS) which is co-administered by the Centers for Disease Control (CDC) and the US Food and Drug Administration (FDA). VAERS reports can be submitted by anyone, but healthcare providers and drug manufacturers are required by law to submit any adverse events they become aware of. This data set is publicly available (vaers.hhs.gov) for viewing or analysis.

With this as background, there are two sub-themes evident in the pharmacovigilance theme, applications of Big Data tools to pharmacovigilance and the methods used in performing these tasks. The applications of Big Data to this important process form a rather large group of papers (65). A

**Table 2. Papers Included in the biology of vaccines theme**

Sub-Theme	Author(s)
Prediction of Biological Processes	<p><b>Epitopes:</b> Kar et al, 2018; Khanna &amp; Rana, 2020; Kim et al, 2019; Sher, Zhi &amp; Zhang, 2017; Solihah, Azhari &amp; Musdholifah, 2020; Rosyda, Adji &amp; Setiawan, 2016; Zhang &amp; Niu, 2010</p> <p><b>Proteins:</b> El-Manzalawy et al, 2016; Munteanu et al, 2014; Nanni &amp; Lumini, 2009; Singh, Singh &amp; Sisodia, 2019</p> <p><b>Peptides:</b> Boehm et al, 2019; Degoot, Chirove &amp; Ndifon, 2018; Henneges, Huster &amp; Zell, 2009; Liu et al 2020; Vang &amp; Xie, 2017</p> <p><b>Antigen/Anitbody:</b> Conti &amp; Karplus, 2019; Khanna &amp; Rana, 2017; Kim et al, 2018; Nagpal et al, 2018; Smith et al, 2019</p> <p><b>Vaccine Effectiveness:</b> Cotugno et al, 2020; Hemedan et al, 2020; Lee et al, 2015; Parvenderh et al, 2019</p>
Description & Analysis	Kwofie et al, 2019; Miranda et al, 2018; Tumiotto et al, 2017; Young et al, 2005
Vaccine Responses	Chaudhury et al, 2020; Flanagan, Noho-Konteh & Dickinson, 2013; Pittala, Morrison & Ackerman, 2019; Tomic et al, 2019
Overviews & Reviews	Cotugno et al, 2015; Izak, Kim & Kaczanowski, 2018; Olafsdottir, Lindqvist & Harandi, 2015
Reverse Vaccinology	Dalsass et al, 2019; Heinson et al, 2019; Ong et al, 2020

sample of these studies is characterized by studies that determine the safety of a particular vaccine (Myers et al 2017; Haber et al, 2019; Haber et al, 2020). Other topics of study within this sub-theme include studies that attempt to develop a profile of those who might be at risk with a particular immunization, such as a particular pneumococcal vaccine (Yildirim et al 2020; Iqbal et al 2015), and studies that seek to link vaccine benefits with ambient conditions such as air pollution (Liu et al, 2020).

The second sub-theme discusses various methods used to perform the pharmacovigilance process. Many of these authors describe improved methods of sharing data, such as Briggs' (2017) study on the Standard for Exchange of Nonclinical Data (SEND). Others support the use of ontologies as a means to improve the usage of data across disciplines and nations (He, 2014). There are also authors who describe various data sources that can aid in pharmacovigilance such as the Influenza Research Database (Zhang et al, 2017; Trifiro, Sultana & Bate, 2018). Lastly, a number of researchers provide methodological improvements for mining these databases such as the effects of timing (Berendsen et al, 2016; Svanström, Callréus & Hviid, 2010; Cai et al, 2017), sample stratification (Woo et al, 2008), and subgroup analysis (de Bie et al, 2012). Table 3 summarizes the papers in the Pharmacovigilance theme.

### *Theme: Public Sentiment About Vaccines*

The only way in which vaccines can be effective is to get them into the arms of the people they are meant to protect. In some cases, these immunizations are required by law, as in the case of children in the US who must be inoculated against a host of childhood diseases before attending school (though there exist many exemptions from this requirement). However, many vaccines are administered only when the patient consents. In some cases, public sentiment against the use of vaccines can thwart the effectiveness of vaccines because of a low acceptance rate. Whether the lack of trust is based on fear, misinformation, or other factors, the ability of a pharmaceutical firm, and in some cases the government, to engender trust in the public is the key to a successful inoculation program. This theme includes those studies that seek to understand the role of public sentiment in vaccine usage and how this public sentiment is influenced.

This theme consists of two sub-themes, which are closely related. The first sub-theme concerns the role of social media in the shaping and movement of public sentiment both for and against vaccines. Certainly, the largest number of these studies focuses on the use of the Twitter platform as a communication and influencing tool. Studies that seek to understand the role of tweets in the decision to take a specific vaccine make up the largest portion of this group, with studies on the usage of influenza (Krittanawong et al, 2017; Kagashe, Yan & Suheryani, 2017; Chan, Jamieson & Albarricin, 2020) human papillomavirus (Du et al, 2017; Massey et al, 2016; Dunn et al, 2017), and the measles-mumps-rubella vaccine (Pananos et al, 2017) leading the field. Other studies in this sub-theme center on the relationship between Twitter and the overarching public opinion on vaccinations (Tavoschi et al, 2020), partisan politics' impact on vaccinations (Walter, Ophir & Jamieson, 2020), vaccines during pregnancy (Martin et al, 2020), and vaccination-related autism (Tomeny, Vargo & El-Toukhy, 2017).

**Table 3. Papers included in the pharmacovigilance theme**

Sub-Theme	Author(s)
Applications of Big Data to Pharmacovigilance	Haber et al, 2020; Haber et al, 2019; Iqbal et al, 2015; Liu et al, 2020; Myers et al, 2017; Yildirim et al, 2020
Methodological	Berendsen et al, 2016; Briggs, 2017; Cai et al, 2017; de Bie et al, 2012; He, 2014; Trifiro, Sultana & Bate, 2018; Svanström, Callréus & Hviid, 2010; Woo et al, 2008; Zhang et al, 2017

Not all studies center on Twitter. The content generated on other websites is also of concern. Getman et al (2018) study the effect of information from other sources such as Wikipedia and the New York Times on vaccine hesitancy, Japanese researchers study the impact of large Japanese datasets on vaccine hesitancy (Nawa et al, 2016; Okuhara et al, 2018; Okuhara et al, 2019a) and Meyer et al (2019) used the Canadian Broadcasting Corp. website as a target. Some researchers also look at the impact of newspapers on the vaccination decision (Okuhara et al, 2019b). Other researchers examine social media as a tool to promote on-time vaccinations for children (Chandir et al, 2018; Bell et al, 2019).

The second sub-theme takes as its target the use of bots in the creation of social media posts to influence the vaccination decision. Yuan, Schuchard & Crooks (2019) study the communication patterns of pro- and anti-vaccine users and the role of bots in shaping those patterns. Kudugunta & Ferrara (2018) study the problem of detecting bots so that they might be identified and their content flagged or removed from the platform. Table 4 contains the papers contained in this theme.

### ***Theme: Technology of Vaccinology***

The specific technologies involved in vaccinology are the subject of this theme. The first sub-theme covers the various overviews and reviews that are prevalent in these papers. These reviews vary in specificity from taking a very wide view to something very narrow, but in all cases they represent a very useful resource for gaining an understanding of the role of information technology, and Big Data tools specifically, in the development of vaccines. de Sousa, de Menezes Neto, and de Brito (2013) explain the challenges of mining the data generated by the DNA sequencing processes used to predict protein behavior. Oberg and colleagues (2015) present some of the lessons learned in dealing with high-dimensional, high-throughput data as applied to vaccinomics. Rodrigo and Luciana (2019) discuss the applications of next generation sequencing (NGS) in studying host-pathogen interactions. Finally, and aptly, Vaishya et al (2020) discuss the role of artificial intelligence in the fight against Covid-19.

Databases play a huge role in computer-aided vaccinology. Because the huge amounts of data generated by tools like NGS need to be formatted in a specific way to be ready for various types of analyses, the database technologies utilized can make a significant impact on system performance (Davies et al, 2007). Tomic et al describe the FluPRINT dataset to enable large studies that explore the underlying concepts of the influenza virus from a cellular and molecular level. To bridge the gap between the immunology and ML communities, the Dana-Farber Cancer Institute has developed a tool to provide standardized datasets to enhance the compatibility of ML-based research in virology and immunology. Finally, the National Institutes of Health (NIH) has developed the NIH Immune Epitope Database, which contains curated datasets created to enhance the prediction of peptide/MHC binding.

Finally, the third sub-theme includes papers that describe certain architectural aspects of systems built specifically for reverse vaccinology research. For example, Dharayani and colleagues (2019) describe a MapReduce-based architecture to run the BLAST Algorithm, an open source tool for

**Table 4. Papers included in the public sentiment theme**

Sub-Theme	Author(s)
Social Media	Chan, Jamieson & Albarracin, 2020; Du et al, 2017; Dunn et al, 2017; Getman et al, 2018; Kagashe et al, 2017; Krittanawong et al, 2017; Massey et al, 2016; Martin et al, 2020; Meyer et al, 2019; Nawa et al, 2016; Okuhara et al, 2019a,b; Okuhara et al, 2018; Pananos et al, 2017; Tavoichi et al, 2020; Tomeny, Vargo & El-Toukhy, 2017; Walter, Ophir & Jamieson, 2020
Bots	Bell et al, 2019; Chandir et al, 2018; Kudugunta & Ferrara, 2018; Yuan, Schuchard & Crooks, 2019

comparing biological sequence information. Table 5 displays some of the papers included in the Technology of Vaccinology theme.

### *Theme: Clinical Trials*

As the world has witnessed over the past year, clinical trials are at the heart of vaccine evaluation and approval. Without data from a sizable group of volunteers who have received the vaccine, there would be no basis for the approval and distribution of a vaccine, regardless of how well it performs in the lab or in a computational simulation. In this theme, the literature describes the use of Big Data technologies in both the pre-trial and post-trial phases of the process. In terms of pre-trial uses, Big Data tools have found their way into the process of matching patients to study acceptance criteria (Henderson, 2016), supporting sequential patient recruitment, sometimes a result of paused trials (Zame et al, 2020), and using deep learning tools to simplify the process of understanding eligibility criteria (Chuan, 2018). Moreover, Morgan (2019) uses analytical tools to pinpoint, and reduce, the costs of clinical trials, which are more expensive to run than most people expect.

During and after the clinical trial is complete, there are many opportunities to utilize Big Data technologies. Lopalco and DeStefano (2015) use datamining techniques to evaluate both Phase 3 trial findings and post-licensure pharmacovigilance to evaluate the safety and efficacy of a vaccine. Machine learning tools are useful when evaluating the actual protection value of the chicken pox and shingles vaccines (Gilbert & Luedtke, 2018; Ackerman, Barouch & Alter, 2017). Large database access and mining tools make analysis of post-trial data easier and more transparent, as well as making the results of these analyses more useful (Zung, 2019). The papers included in the Clinical Trials theme are displayed in Table 6.

### *Theme: Miscellaneous*

This theme consists of Big Data tools that are used beyond those activities outlined earlier. These applications, though not numerous enough to be given the status of theme, nevertheless are an important part of the vaccine process and should be enumerated. Within this theme, there are a number of interesting areas of research. Agarwal et al (2020) demonstrate the usefulness of deep learning processes to develop a profit function for the pharmaceutical firm. The manufacturing of vaccines is the focus of a number of efforts. Wu et al (2019) use a ML approach to perform pose estimation to

**Table 5. Papers included in the technology of vaccinology theme**

Sub-Theme	Author(s)
Overview	Banerjee, Basu & Nasipuri, 2015; Bragazzi et al, 2018; de Sousa, De Menzes Neto & De Brito, 2013; Lalmuanawma, Hussain & Chhakchuak; Oberg et al, 2015; Rodrigo & Luciani, 2019; Vaishya et al, 2020
Databases	Altmann, 2018; Tomic et al, 2019; Zhang et al, 2011
Architecture	Dharayani et al, 2019

**Table 6. Papers included in the clinical trials theme**

Sub-Theme	Author(s)
Pre-Trial	Chuan, 2018; Henderson, 2016; Morgan, 2019; Zame et al, 2020
Post-Trial	Ackerman, Barouch & Alter, 2017; Gilbert & Luedtke, 2018; Lopalco & DeStefano, 2015; Zung, 2019



support fully automatic mosquito salivary gland extraction, a step necessary to creation of malaria vaccine. Neural network tools are used by Yu et al (2019) to determine the fertility of chicken embryos, important for the manufacture of a variety of vaccines.

The sophisticated supply chain necessary to maintain the proper conditions while vaccines are distributed, usually focused on the temperature of the vaccine, benefits from the use of Big Data tools. Yong et al (2020) suggest a blockchain-based approach to vaccine supply chain management using ML tools to analyze the chain’s performance. Sujaree (2019) suggests the use of the max-min ant system (MMAS) algorithm to design a vaccine cold chain. Bhattacharjee et al (2015) develop a system to integrate the various sources of data generated during the transportation of vaccines in an effort to provide a more comprehensive portrait of the transportation process.

The topic of immunization protocols is addressed by researchers using Big Data tools. Chen et al (2020) use ML to develop clinical decision support system for shingles vaccination and Hovav et al (2017) take a similar approach using health care analytics to support an influenza vaccination program. Finally, Bhatti et al (2018) utilize datamining to help create a recommender system to support optimal coverage of vaccinations and Clark et al (2011) propose the use of patent data mining to find vaccines that can be repurposed for other uses to avoid the time and expense of new vaccine development. Table 7 presents the papers categorized in the Miscellaneous theme.

DISCUSSION

The results of this study illustrate the wide variety of uses to which Big Data tools have been put in the development and deployment of vaccines. These tools have been used to quicken the development of vaccines as well as making them more effective, safer, and less costly. Within these results, there are a number of interesting trends and issues that necessitate further discussion. It is apparent by the preponderance of the literature that the use of these tools within the biological activities is well accepted and a very appropriate application. These tools are designed for the purposes of analyzing combinations of many variables such as the combination of epitopes and antibodies. Likewise, the second largest piece of the literature centers on the use of datamining technologies to evaluate the huge databases that contain data on adverse outcomes from vaccine use. With the other themes being much smaller, it seems that the research world has either not seen the need for these tools in these various areas or that they have not yet addressed them.

As a tool to understand and plan Big Data implementations, there have been many “stacks” put forth by authors across the Big Data and analytics landscape. These stacks are frameworks developed to provide a convenient roadmap for researchers, students, and practitioners to follow when designing, analyzing, or learning about Big Data implementations and how they can be deployed. For example, Frampton (2018) describes a Spark-based stack for the creation of an open source Big Data stack, Lu et al (2018) present a Deep Learning Over Big Data (DLoBD) stack that provides a more specific approach to combining Big Data with high performance computing and compare performance across multiple configurations, and Sakr (2017) analyzes the performance variations between Apache and Hadoop – based stacks. For the present study, a Big Data stack might be a valuable lens through which

Table 7. Papers Included in the miscellaneous theme

	Author(s)
Miscellaneous Theme	Agarwal et al, 2020; Bhattacharjee et al, 2015; Bhatti et al, 2018; Chen et al, 2020; Clark et al, 2011; Hovav et al, 2017; Montagna et al, 2020; Pennisi, Russo & Pappalardo, 2018; Sujaree, 2019; Wu et al, 2019; Yong et al, 2020; Yu et al, 2019



to view the contributions of some of the authors. While it is infeasible to evaluate each contribution through this lens, a few examples of each type will be useful. The simplified stack in use here consists of four layers:

1. **Data Storage:** technologies that store huge masses of data such as Hadoop or Amazon S3.
2. **Integration & Ingestion:** technologies that perform the ingestion and management of data from their sources (Ex. Stitch, Apache Kafka) as well as provide Extract, Transform, & Loading (ETL) services (Ex. Python).
3. **Processing:** technologies to perform the calculations necessary to address the research questions (Ex. Apache Spark, Map/Reduce, Tensorflow).
4. **Analytics:** technologies to perform analytical calculations as well as provide visual analytics and dashboards (Ex. Tableau, R).

With this framework in place, Table 8 provides some examples of tools in these layers used to improve vaccinology. It should be noted that it is unrealistic to describe each study as only using one of these layers of technology. In reality, each portion of the stack is being used in almost every effort. For example, when an author describes the use of a machine learning tool to predict the binding propensity of a peptide, it is assumed that they are also using some form of data storage tool, probably employing ETL or other pipelining method to improve the flow, and analyzing results on a

**Table 8. Technology application with reference to big data stack**

Layer	Author	Tool	Results
1. Data Storage	Zhang et al, 2016 Miranda et al 2018 Massey et al, 2016	Influenza Research Database Kröhnke Pyridine Library multiple data collection tools	Describe use of DB to enhance research collaboration Demonstrate value of DB in research of arenavirus Analyze Twitter data to quantify HPV communications
2. Integration & Ingestion	Briggs, 2016	Standard Exchange for Nonclinical Data	Describe use of common data format to reduce the need for certain ETL activities
3. Processing	Dharayani et al, 2019 Flanagan et al, 2013 Degoot, Chirove & Ndifon, 2018	BLAST algorithm on Map/Reduce platform datamining tools machine learning	Decreased time required to search for genetic anomalies Improve study of vaccine response data One of many papers using ML to predict peptide or epitope interactions
4 Analytics	Woo et al, 2008 Dunn et al, 2017 Chandir et al, 2018	visual analytics GIS, visual analytics predictive analytics	Study the effects of sample stratification when datamining vaccine responses Determine impact of info exposure on HPV vaccine coverage Identify children at high risk of missing vaccinations

visualization or analytic platform. The table, and this study, classifies papers based on the technology highlighted by the author in terms of their perceived contributions.

The four layers of the Big Data stack are highlighted in Table 8 as a means of further illuminating the value of these tools in the development of vaccines. Those authors noted in the Data Layer have utilized some form of large dataset as a tool to leverage further research and spur communication and collaboration among vaccine researchers. Without these standardized datasets, the complexity of sharing data is greatly complicated because of differences in format, labeling, and simple access. This leads to the second layer, Integration and Ingestion. This layer could also include tools such as data warehousing and the like. While these tools are likely used by most of the researchers in the study, they were only noted by one who described the use of the SEND format when sharing data across institutions or projects. Using a standard data format removes many of the typical ETL tasks involved in making data from external sources useful to a project.

The use of various machine learning techniques is common among those researchers trying to predict certain biological aspects of vaccinology but a full comparison of the various

tools and their performance is beyond the scope of this paper. Even though some of the papers focused on comparing the performance of specific tools on a single task, the tasks varied across a wide spectrum of size and computations, making more global comparisons difficult. However, researchers such as Rosyda, Adji & Setiawan (2016) who use a neural network approach to predicting epitope activity on the P24 protein of the HIV virus provides an example of how these tools are employed. The amino acids within the epitope are first encoded and then used to create a training set for use in developing the neural network capabilities. These researchers found that multiple training sessions, known as cascade training, provided even better predictive performance and sensitivity.

The same can be said about the use of datamining tools in the performance of pharmacovigilance studies. These tools provide an important window into the performance of a vaccine after approval and implementation, but the specifics of each study differ widely. Another example of how datamining is put to use these tools to analyze data collected from throughout a human body to describe its response to a vaccination (Flanagan et al, 2013). These huge and complex datasets can be mined to allow researchers to understand on a much broader scale the human body's response to vaccinations.

The last category, Analytics, comprises all of the tools that researchers might use to analyze the results of their computation or try to draw information out of large databases, such as those described in the first layer of the stack. In most cases, the actual tool being used is not described fully, only the functionalities employed. In many cases, visual analytics are employed to understand patterns and relationships in the data or to enable interactions hidden from view to be highlighted in a useful manner. An illustration of this approach is the use of visual analytics and Geographical Information Systems (GIS) to determine the regional differences in vaccine acceptance based on social media data (Dunn et al, 2017). Using GIS tools, the data are not changed or transformed, but displayed in such a way that the underlying effects of location can be highlighted to enable easier understanding of the differences brought about by that single variable.

There is a large disparity between the number of journal articles and conference proceedings. This was surprising because many times, if not most, when information technologies are introduced into a new milieu, the initial forays into that area of inquiry happen at conferences. These conferences are where new ideas are tried out and then published in journals after they mature. In the present case, journal articles are much more numerous than conference papers. This might be an artifact of the manner in which the databases are searched or constructed, or possibly the nature of the industry in which these efforts take place. It would be interesting to see if other areas of literature in similar circumstances exhibit the same characteristics.

Another interesting characteristic of this body of literature is that, unlike other areas of technological inquiry, there are many more empirical papers than theoretical. Again, in other technological areas of inquiry, early literature is theoretical and/or prescriptive in nature, and then become more applied as the field matures (Kasten, 2020). In this case, there are many more empirical

uses of these technologies than prescriptive descriptions of systems yet to be built. Part of the change of balance has to do with the huge chunk of literature that deals specifically with analysis of adverse events surveillance. But, even when that literature is taken out, the large proportion of applied and empirical studies suggest that the researchers in this area have been better able to apply these technologies. Perhaps this represents a shortening of the learning curve as big data tools become more widely used and understood.

Very surprising is the lack of research surrounding the application of these technologies to the topics contained in the Miscellaneous theme such as supply chain management and manufacturing. It is hard to believe that the difficulties associated with transporting highly unstable materials such as vaccines, with their formidable temperature requirements, would not benefit from additional analytical tools and resulting information. The same might be said of the manufacturing function, which include very complicated processes and materials. The possibility exists that this research has taken place but is not narrowly identified as vaccine related literature, but rather more broadly in the realm of pharmaceuticals. Future research will explore this possibility.

Structured literature reviews have their inherent limitations. Choices of databases, search terms, or combinations thereof, can have significant effects on the results. This is coupled with the requirement that the researcher must exert judgement over which category or theme a particular document belongs to. Thus, differences in opinion over a particular document might result in slightly different results for a specific researcher. There is also the difficulty that classifying documents into closely related sub-themes can be very difficult because they deal with topics that are tightly interconnected with overlapping issues. For this study, each document was analyzed to determine what the most likely emphasis of the author is. There are many that could have found a home in more than one category and that is the very nature of interdisciplinary research. The difficulty in classification is exactly the reason for the value brought by this type of research.

## CONCLUSION

The current study is intended to create an understanding in the reader of the uses to which Big Data tools such as analytics, datamining, machine learning and others within the vaccine development and distribution industry. The study also seeks to provide a framework for researchers to perform further research, especially in those areas which, as demonstrated by this study, have yet to be thoroughly investigated. Further research in this field is required to continue to illuminate the topic of Big Data technology in the vaccine industry specifically, and the pharmaceutical industry generally. While this would be helpful to researchers, it would also be helpful to practitioners and clinicians who are narrowly concerned with a specific aspect of the process and might not have a broader concept of the tools and technologies available to them.

**REFERENCES**

- Ackerman, M. E., Barouch, D. H., & Alter, G. (2017). Systems serology for evaluation of HIV vaccine trials. *Immunological Reviews*, 275(1), 262–270. doi:10.1111/immr.12503 PMID:28133810
- Agarwal, P., Tamer, M., Sahraei, H., & Budman, H. (2020). Deep learning for classification of profit-based operating regions in industrial processes. *Industrial & Engineering Chemistry Research*, 59(6), 2378–2395. doi:10.1021/acs.iecr.9b04737
- Altman, D. M. (2018). New tools for MHC research from machine learning and predictive algorithms to the tumour immunopeptidome. *Immunology*, 154(3), 329–330. doi:10.1111/imm.12956 PMID:29902342
- Banerjee, S., Basu, S., & Nasipuri, M. (2015). Big Data analytics and its prospects in computational proteomics. In *Information Systems Design and Intelligent Application* (pp. 591-598). Springer. doi:10.1007/978-81-322-2247-7\_60
- Bell, A., Rich, A., Teng, M., Oreskovic, T., Bras, N. B., Mestrinho, L., Golubovic, S., Pristas, I., & Zejnilovic, L. (2019). Proactive advising: A machine learning driven approach to vaccine hesitancy. *IEEE International Conference on Healthcare Informatics*.
- Berendsen, M. L. T., Smits, J., Netea, M. G., & van der Ven, A. (2016). Non-specific effects of vaccines and stunting: Timing may be essential. *EBioMedicine*, 8, 341–348. doi:10.1016/j.ebiom.2016.05.010
- Bhattacharjee, P. S., Solanki, M., Bhattacharyya, R., Ehrenberg, I., & Sarma, S. (2015). VacSeen: A linked data-based information architecture to track vaccines using barcode scan authentication. SWAT4LS.
- Bhatti, U. A., Huang, M., Wang, H., Zhang, Y., Mehmood, A., & Di, W. (2018). Recommendation system for immunization coverage and monitoring. *Human Vaccines & Immunotherapeutics*, 14(1), 165–171. doi:10.1080/21645515.1379639
- Boehm, K. M., Bhinder, B., Raja, V. J., Dephoure, N., & Elemento, O. (2019). Predicting peptide presentation by major histocompatibility complex class I: An improved machine learning approach to the immunopeptidome. *BMC Bioinformatics*, 20(7), 1–11. doi:10.1186/s12859-018-2561-z PMID:30611210
- Bragazzi, N. L., Gianfredi, V., Villarini, M., Rosselli, R., Nasr, A., Hussein, A., Martini, M., & Behzadifar, M. (2018). Vaccines meet Big Data: State-of-the-art and future prospects. From the classical 3Is (“Isolate-Inactivate\_Inject”) Vaccinology 1.0 to Vaccinology 3.0 Vaccinomics, and beyond: A historical overview. *Frontiers in Public Health*, 6(62), 1–9. doi:10.3389/fpubh.2018.00062 PMID:29556492
- Briggs, K. (2017). Making sense of SEND: The standard for exchange of nonclinical data. *Regulatory Toxicology and Pharmacology*, 91, 77–85. doi:10.1016/j.yrtph.2017.10.012 PMID:29066334
- Briner, R. B., & Denyer, D. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. In D. M. Rousseau (Ed.), *The Oxford Handbook of Evidence-Based Management*. Oxford University Press.
- Cai, Y., Du, J., Huang, J., Ellenberg, S. S., Hennesy, S., Tao, C., & Chen, Y. (2017). A signal detection method for temporal variation of adverse effect with vaccine adverse event reporting system data. *BMC Medical Informatics and Decision Making*, 17(76), 93–100. doi:10.1186/s12911-017-0472-y PMID:28699543
- Chan, M.-S., Jamieson, K. H., & Albarracin, D. (2020). Prospective associations of regional social media messages with attitudes and actual vaccination: A big data and survey study of the influenza vaccine in the United States. *Vaccine*, 38(40), 6236–6247. doi:10.1016/j.vaccine.2020.07.054 PMID:32792251
- Chandir, S., Siddiqi, D. A., Hussain, O. A., Naizi, T., Shah, M. T., Dharma, V. K., Habib, A., & Khan, A. J. (2018). Using predictive analytics to identify children at high risk of defaulting from a routine immunization program: Feasibility study. *JMIR Public Health and Surveillance*, 4(3), 1-12.
- Chaudhury, S., Duncan, E. H., Atre, T., Dutta, S., Spring, M. D., Leitner, W. W., & Bergmann-Leitner, E. S. (2020). Combining immunoprofiling with machine learning to assess the effects of adjuvant formulation on human vaccine-induced immunity. *Human Vaccines & Immunotherapeutics*, 16(2), 400–411. doi:10.1080/21645515.2019.1654807 PMID:31589550

- Chen, J., Chokshi, S., Hegde, R., Gonzalez, J., Iturrate, E., Aphinyanaphongs, Y., & Mann, D. (2020). Development, implementation, and evaluation of a personalized machine learning algorithm for clinical decision support: Case study with shingles vaccination. *Journal of Medical Internet Research*, 22(4), 1–12. doi:10.2196/16848 PMID:32347813
- Chuan, C.-H. (2018). Classifying eligibility criteria in clinical trials using active deep learning. *17th IEEE International Conference on Machine Learning and Applications*.
- Clark, K., Cavicchi, J., Jensen, K., Fitzgerald, R., Bennett, A., & Kowalski, S. P. (2011). Patent data mining: A tool for accelerating HIV vaccine innovation. *Vaccine*, 29(24), 4086–4093. doi:10.1016/j.vaccine.2011.03.052
- Conti, S., & Karplus, M. (2019). Estimation of the breadth of CD4bs targeting HIV antibodies by molecular modeling and machine learning. *PLoS Computational Biology*, 15(4), 1–22. 10.1371/journal.pcbi.1006954
- Cotugno, N., De Armas, L., Pallikkuth, S., Rossi, P., Palma, P., & Pahwa, S. (2015). Paediatric HIV infection in the ‘omics era: Defining transcriptional signatures of viral control and vaccine responses. *Journal of Virus Eradication*, 1(3), 153–158. doi:10.1016/S2055-6640(20)30507-0 PMID:26807446
- Cotugno, N., Santilli, V., Pascucci, G. R., Manno, E. C., De Armas, L., Pallikkuth, S., Deodati, A., Amodio, D., Zangari, P., Zicari, S., Ruggiero, A., Fortin, M., Bromley, C., Pahwa, R., Rossi, P., Pahwa, S., & Palma, P. (2020). Artificial intelligence applied to *in vitro* gene expression testing (IVIGET) to predict trivalent inactivated influenza vaccine immunogenicity in HIV infected children. *Frontiers in Immunology*, 11, 1–13. doi:10.3389/fimmu.2020.559590 PMID:33123133
- Dalsass, M., Brozzi, A., Medini, D., & Rappuoli, R. (2019). Comparison of open-source reverse vaccinology programs for bacterial vaccine antigen discovery. *Frontiers in Immunology*, 10, 1–12. Advance online publication. doi:10.3389/fimmu.2019.00113 PMID:30837982
- de Bie, S., Verhamme, K. M. C., Straus, S. M. J. M., Stricker, B. H. C., & Sturkenboom, M. C. J. M. (2012). Vaccine-based subgroup analysis in VigiBase. *Drug Safety*, 35(4), 335–346. doi:10.2165/11598120-000000000-00000 PMID:22435344
- de Sousa, T. N., & de Menezes Neto, A. (2013). “Omics” in the study of the major parasitic diseases malaria and schistosomiasis. *Infection, Genetics and Evolution*, 19, 258–273. doi:10.1016/j.meegid.2013.07.008 PMID:23871773
- Degoot, A. M., Chirove, F., & Ndifon, W. (2018). Trans-allelic model for prediction of peptide: MHC-11 interactions. *Frontiers in Immunology*, 9, 1–11. Article, 1410. Advance online publication. doi:10.3389/fimmu.2018.01410 PMID:29988560
- Dharayani, R., Wibowo, W. C., Ruldeviyani, Y., & Gandhi, A. (2019). *Genomic anomaly searching with BLAST algorithm using MapReduce framework in big data platform*. IWBIS.
- Du, J., Xu, J., Song, H., & Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making*, 17(69), 63–70. doi:10.1186/s12911-017-0469-6
- Dunn, A. G., Surian, D., Leask, J., Dey, A., Mandl, K. D., & Coiera, E. (2017). Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. *Vaccine*, 35(23), 3033–3040. doi:10.1016/j.vaccine.2017.04.060 PMID:28461067
- El-Manzalawy, Y., Munoz, E. E., Lidner, S. E., & Honavar, V. (2016). PlasmoSEP: Predicting surface-exposed proteins on the malaria parasite using semisupervised self-training and expert-annotated data. *Proteomics*, 16(23), 2967–2976. doi:10.1002/pmic.201600249 PMID:27714937
- Flanagan, K. L., Noho-Konteh, F., Ghazal, P., & Dickinson, P. (2013). Transcriptional profiling technology for studying vaccine responses: An untapped goldmine. *Methods (San Diego, Calif.)*, 60(3), 269–274. doi:10.1016/j.ymeth.2013.03.032 PMID:23578546
- Frampton, M. (2018). *Complete Guide to Open Source Big Data Stack*. Springer. doi:10.1007/978-1-4842-2149-5
- Getman, R., Helmi, M., Roberts, H., Yansane, A., Cutler, D., & Seymour, B. (2018). Vaccine hesitancy and online information: The influence of digital networks. *Health Education & Behavior*, 45(4), 599–606. doi:10.1177/1090198117739673 PMID:29267129

Gilbert, P. B., & Luedtke, A. R. (2018). Statistical learning methods to determine immune correlates of herpes zoster in vaccine efficacy trials. *The Journal of Infectious Diseases*, 218(suppl\_2), s99–s101. doi:10.1093/infdis/jiy421 PMID:30247601

Haber, P., Moro, P. L., Ng, C., Doros, G. M., Perez-Vilar, S., Marquez, P. L., & Cano, M. (2020). Safety review of tetanus toxoid, reduced diphtheria toxoid, acellular pertussis vaccines (Tdao) in adults aged  $\geq 65$  years, Vaccine Adverse Event Reporting System (VAERS), United States, September 2010–December 2018. *Vaccine*, 38(6), 1476–1480. doi:10.1016/j.vaccine.2019.11.074 PMID:31883809

Haber, P., Moro, P. L., Ng, C., Doros, G. M., Lewis, P., & Cano, M. (2019). Post licensure surveillance of trivalent adjuvanted influenza vaccine (allV3;Fluad), vaccine Adverse Event Reporting System (VAERS), United States, July 2016–June 2018. *Vaccine*, 37(11), 1516–1520. doi:10.1016/j.vaccine.2019.01.052 PMID:30739795

He, Y. (2014). Ontology-supported research on vaccine efficacy, safety and integrative biological networks. *Expert Review of Vaccines*, 13(7), 825–841. doi:10.1586/14760584.2014.923762 PMID:24909153

Heinson, A. I., Ewing, R. M., Holloway, J. W., Woelk, C. H., & Niranjani, M. (2019). An evaluation of different classification algorithms for protein sequence-based reverse vaccinology prediction. *PLoS One*, 14(12), 1–13. doi:10.1371/journal.pone.0226256 PMID:31834914

Hemedan, A., Elaziz, M. A., Jiao, P., Alavi, A. H., Bahgat, M., Ostaszewski, M., Schneider, R., Ghazy, H. A., Ewees, A. A., & Lu, S. (2020). Prediction of the vaccine-derived Poliovirus outbreak incidence: A hybrid machine learning approach. *Scientific Reports*, 10(5058), 1–12. doi:10.1038/s41598-020-61853-y PMID:32193487

Henderson, L. (2016). Innovations in patient matching. *Applied Clinical Trials*, 25(8/9), 1.

Henneges, C., Huster, S., & Zell, A. (2009). An artificial T cell immune system for predicting MHC-II binding peptides. *IEEE Symposium on Artificial Life*.

Hovav, S., Tell, H., Levner, E., Ptuskin, A., & Herbon, A. (2017). Health care analytics and Big Data management in influenza vaccination programs: Use of information-entropy approach. In *The Analytics Process* (pp. 211–237). doi:10.1109/ALIFE.2009.4937708

Iqbal, S., Shi, J., Seib, K., Lewis, P., Moro, P. L., Woo, E. J., Shimabukuro, T., & Orenstein, W. A. (2015). Preparation for global introduction of inactivated poliovirus vaccine: Safety evidence from the US Vaccine Adverse Event Reporting System, 2000–12. *The Lancet. Infectious Diseases*, 15(10), 1175–1182. doi:10.1016/S1473-3099(15)00059-6 PMID:26289956

Izak, D., Klim, J., & Kaczanowski, S. (2018). Host-parasite interactions and ecology of the malaria parasite-a bioinformatics approach. *Briefings in Functional Genomics*, 17(6), 451–457. doi:10.1093/bfpg/ely013 PMID:29697785

Jordan, K., Dossou, P.-E., & Chang, J. Jr. (2019). Using lean manufacturing and machine learning for improving medicines procurement and dispatching in a hospital. *Procedia Manufacturing*, 38, 1034–1041. doi:10.1016/j.promfg.2020.01.189

Kagashe, I., Yan, Z., & Suheryani, I. (2017). Enhancing seasonal influenza surveillance: Topic analysis of widely used medicinal drugs using Twitter data. *Journal of Medical Internet Research*, 19(9), 1–14. doi:10.2196/jmir.7393 PMID:28899847

Kar, P., Ruiz-Perez, L., Arooj, M., & Mancera, R. L. (2018). Current methods for the prediction of T-cell epitopes. *Peptide Science (Hoboken, N.J.)*, 110(2), 1–17. doi:10.1002/pep2.24046

Kasten, J. (2020). Big data applications in healthcare administration. *International Journal of Big Data Applications in Healthcare*, 5(2), 12–37.

Khanna, D., & Rana, P. S. (2017). Multilevel ensemble model for prediction of IgA and IgG antibodies. *Immunology Letters*, 184, 51–60. doi:10.1016/j.imlet.2017.01.017 PMID:28214535

Khanna, D., & Rana, P. S. (2020). Improvement in prediction of antigenic epitopes using stacked generalisation: An ensemble approach. *IEEE Systems Biology*, 14(1), 1–7. doi:10.1049/iet-syb.2018.5083 PMID:31931475

- Kim, S., Kim, H. S., Kim, E., Lee, M. G., Shin, E.-C., Paik, S., & Kim, S. (2018). Neopepsee: Accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 29(4), 1030–1036. doi:10.1093/annonc/mdy022 PMID:29360924
- Kim, Y., Lee, J., Ha, K., Lee, W.-K., Heo, D. R., Woo, J. R., & Yu, S. J. (2019). A computational framework for deep learning-based epitope prediction by using structure and sequence information. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.
- Krittana Wong, C., Tunhasiriwet, A., Chirapongsathorn, S., & Kitai, T. (2017). Tweeting influenza vaccine to cardiovascular health community. *European Journal of Cardiovascular Nursing*, 16(8), 704–706. doi:10.1177/1474515117707867
- Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, 467, 312–322. doi:10.1016/j.ins.2018.08.019
- Kwofie, S. K., Broni, E., Teye, J., Quansah, E., Issah, I., Wilson, M. D., Miller, W. A. III, Tiburu, E. K., & Bonney, J. H. K. (2019). Pharmacoinformatics-based identification of potential bioactive compounds against Ebola virus protein VP24. *Computers in Biology and Medicine*, 113, 1–18. doi:10.1016/j.combiomed.2019.103414 PMID:31536833
- Lalmuanawma, S., Hussain, J., & Chhakhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons, and Fractals*, 139, 1–6. doi:10.1016/j.chaos.2020.110059 PMID:32834612
- Lee, A. J., Das, S. R., Wang, W., Fitzgerald, T., Pickett, B. E., Aevermann, B. D., Topham, D. J., Falsey, A. R., & Scheuermann, R. H. (2015). Diversifying selection analysis predicts antigenic evolution of 2009 pandemic H1N1 influenza A virus in humans. *Journal of Virology*, 89(10), 5427–5440. doi:10.1128/JVI.03636-14 PMID:25741011
- Liu, G., Carter, B., Bricken, T., Jain, S., Viard, M., Carrington, M., & Gifford, D. K. (2020). Computationally optimized SARS-CoV-2 MHC class I and II vaccine formulations predicted to target human haplotype distributions. *Cell Systems*, 11(2), 131–144. doi:10.1016/j.cels.2020.06.009 PMID:32721383
- Liu, K., Li, S., Qian, Z. M., Dharmage, S. C., Bloom, M. S., Heinrich, J., Jalaludin, B., Markevych, I., Morawska, L., Knibbs, L. D., Hinyard, L., Xian, H., Liu, S., Line, S., Leskinen, A., Komppula, M., Jalava, P., Roponen, M., Hu, L.-W., Zeng, X.-W., Hu, W., Chen, G., Yang, B.-Y., Guo, Y., & Dong, G.-H. (n.d.). Benefits of influenza vaccination on the associations between ambient air pollution and allergic respiratory diseases in children and adolescents: New insights from the Seven Northeastern Cities study in China. *Environmental Pollution*, 256, 1–10. 10.1016/j.envpol.2019.113434
- Lopalco, P. L., & DeStefano, F. (2015). The complementary roles of phase 3 trials and post-licensure surveillance in the evaluation of new vaccines. *Vaccine*, 33(13), 1541–1548. doi:10.1016/j.vaccine.2014.10.047 PMID:25444788
- Lu, X., Shi, H., Biswas, R., Javed, M. H., & Panda, D. K. (2018). DLoBD: A comprehensive study of deep learning of big data stacks on HPC clusters. *IEEE Transactions on Multi-Scale Computing Systems*, 4(4), 635–648. doi:10.1109/TMSCS.2018.2845886
- Martin, S., Kilich, E., Dada, S., Kummervold, P. E., Denny, C., Paterson, P., & Larsen, H. J. (2020). Vaccines for pregnant women...?!-Mapping maternal vaccination discourse and stance on social media over six months. *Vaccine*, 38(42), 6627–6637. doi:10.1016/j.vaccine.2020.07.072 PMID:32788136
- Massey, P. M., Leader, A., Yom-Tov, E., Budenz, A., Fisher, K., & Klassen, A. C. (2016). Applying multiple data collection tools to quantify human papillomavirus vaccine communication on Twitter. *Journal of Medical Internet Research*, 18(12), 1–11. doi:10.2196/jmir.6670 PMID:27919863
- Meyer, S. B., Violette, R., Aggarwal, R., Simeoni, M., MacDougall, H., & Waite, N. (2019). Vaccine hesitancy and Web 2.0: Exploring how attitudes and beliefs about influenza vaccination are exchanged in online threaded user comments. *Vaccine*, 37(13), 1769–1774. doi:10.1016/j.vaccine.2019.02.028 PMID:30826142
- Miranda, P. O., Cubitt, B., Jacob, N. T., Janda, K. D., & de la Torre, J. C. (2018). Mining a Krohnke Pyridine Library for anti-arenavirus activity. *ACS Infectious Diseases*, 4(5), 815–824. doi:10.1021/acsinfecdis.7b00236 PMID:29405696

# International Journal of Big Data and Analytics in Healthcare

Volume 6 • Issue 2 • July-December 2021

Montagna, M. T., De Giglio, O., Napoli, C., Fasano, F., Diella, G., Donnoli, R., Caggiano, G., Tafuri, S., Lopalco, P. L., & Agodi, A. (2020). Adherence to vaccination policy among public health professionals: Results of national survey in Italy. *Vaccines*, 8(379), 1–16. doi:10.3390/vaccines8030379 PMID:32664507

Morgan, C. (2019). Analytics and metrics help pinpoint costs of study start-up. *Applied Clinical Trials*, 28(1/2), 10–15.

Munteanu, C. R., Pedreira, N., Dorado, J., Pazos, A., Perez-Montoto, L. G., Ubeira, F. M., & Gonzalez-Diaz, H. (2014). LECTINPred: Web server that uses complex networks of protein structure for prediction of lectins with potential use as cancer biomarkers or in parasite vaccine design. *Molecular Informatics*, 33(4), 276–285. doi:10.1002/minf.201300027 PMID:27485774

Myers, T. R., McNeil, M. M., Ng, C. S., Li, R., Lewis, P. W., & Cano, M. (2017). Adverse events following quadrivalent meningococcal CRM-conjugate vaccine (Menveo) reported to the Vaccine Adverse Event Reporting System (VAERS), 2010–2015. *Vaccine*, 35(14), 1758–1763. doi:10.1016/j.vaccine.2017.02.030 PMID:28262331

Nagpal, G., Chaudhary, K., Agrawal, P., & Raghava, P. S. (2018). Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants. *Journal of Translational Medicine*, 16(181), 1–15. doi:10.1186/s12967-018-1560-1 PMID:29970096

Nanni, L., & Lumini, A. (2009). An ensemble of support vector machines for predicting virulent proteins. *Expert Systems with Applications*, 36(4), 7458–7462. doi:10.1016/j.eswa.2008.09.036

Nawa, N., Kogaki, S., Takahashi, K., Ishida, H., Baden, H., Katsuragi, S., Narita, J., Tanaka-Taya, K., & Ozono, K. (2016). Analysis of public concerns about influenza vaccinations by mining a massive online question dataset in Japan. *Vaccine*, 34(27), 3207–3213. doi:10.1016/j.vaccine.2016.01.008 PMID:26776467

Oberg, A. L., McKinney, B. A., Schaid, D. J., Pankratz, V. S., Kennedy, R. B., & Poland, G. A. (2015). Lessons learned in the analysis of high-dimensional data in vaccinomics. *Vaccine*, 33(40), 5262–5270. doi:10.1016/j.vaccine.2015.04.088 PMID:25957070

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future – big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375(13), 1216–1219. doi:10.1056/NEJMp1606181 PMID:27682033

Okuhara, T., Ishikawa, H., Okada, M., Kato, M., & Kiuchi, T. (2018). Contents of Japanese pro- and anti-HPV vaccination websites: A text mining analysis. *Patient Education and Counseling*, 101(3), 406–413. doi:10.1016/j.pec.2017.09.014 PMID:29031425

Okuhara, T., Ishikawa, H., Okada, M., Kato, M., & Kiuchi, T. (2019a). Japanese anti-versus pro-influenza vaccination websites: A test-mining analysis. *Health Promotion International*, 34(3), 552–566. doi:10.1093/heapro/day015 PMID:29584863

Okuhara, T., Ishikawa, H., Okada, M., Kato, M., & Kiuchi, T. (2019b). Newspaper coverage before and after the HPV vaccination crisis began in Japan: A text mining analysis. *BMC Public Health*, 19(770), 1–15. doi:10.1186/s12889-019-7097-2 PMID:31208394

Olafsdottir, T., Lindqvist, M., & Harandi, A. M. (2015). Molecular signatures of vaccine adjuvants. *Vaccine*, 33(40), 5302–5307. doi:10.1016/j.vaccine.2015.04.099 PMID:25989447

Ong, E., Wong, M. U., & He, Y. (2020). COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *Frontiers in Immunology*, 11, 1–13. Advance online publication. doi:10.3389/fimmu.2020.01581 PMID:32719684

Pananos, A. D., Bury, T. M., Wang, C., Schonfeld, J., Mohanty, S. P., Nyhan, B., Salathe, M., & Bauch, C. T. (2017). Critical dynamics in population vaccinating behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 114(52), 13762–13767. Advance online publication. doi:10.1073/pnas.1704093114 PMID:29229821

Parvande, S., Poland, G. A., Kennedy, R. B., & McKinney, B. A. (2019). Multi-level model to predict antibody response to influenza vaccine using gene expression interaction network feature selection. *Microorganisms*, 7(79), 1–17. doi:10.3390/microorganisms7030079 PMID:30875727



- Pennisi, M., Russo, G., & Pappalardo, F. (2018). Combining parallel genetic algorithms and machine learning to improve the research of optimal vaccination protocols. *26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*.
- Pittala, S., Morrison, K. S., & Akerman, M. E. (2019). Systems serology for decoding infection and vaccine-induced antibody responses to HIV-1. *Current Opinions HIV AIDS*, 14, 253–264. doi:10.1109/PDP2018.2018.00070
- Rodrigo, C., & Luciani, F. (2019). Dynamic interactions between RNA viruses and human hosts unravelled by a decade of next generation sequencing. *BBA-General Subjects*, 1863(2), 511–519. doi:10.1016/j.bbagen.2018.12.003 PMID:30528489
- Rosyda, M., Adji, T. B., & Setiawan, N. A. (2016). Cascade neural network for predicting epitope on P24 HIV-1. *Proceedings of the 1st International Conference on Science and Technology 2015 (ICST-2015)*. doi:10.1063/1.4958500
- Sakr, S. (2017). Big data processing stacks. *IT Professional*, 19(1), 34–41. doi:10.1109/MITP.2017.6
- Sher, G., Zhi, D., & Zhang, S. (2017). DRREP: Deep ridge regressed epitope predictor. *BMC Genomics*, 18(S6), 55–65. doi:10.1186/s12864-017-4024-8 PMID:28984193
- Singh, D., Singh, P., & Sisodia, D. S. (2019). Compositional model based on factorial evolution for realizing multi-task learning in bacterial virulent protein prediction. *Artificial Intelligence in Medicine*, 101, 1–6. doi:10.1016/j.artmed.2019.101757 PMID:31813491
- Smith, C. C., Chai, S., Washington, A. R., Lee, S. J., Landoni, E., Field, K., Garness, J., Bixby, L. M., Selitsky, S. R., Parker, J. S., Savoldo, B., Serody, J. S., & Vincent, B. G. (2019). Machine-learning prediction of tumor antigen immunogenicity in the selection of therapeutic epitopes. *Cancer Immunology Research*, 7(10), 1591–1604. doi:10.1158/2326-6066.CIR-19-0155 PMID:31515258
- Solihah, B., Azhari, A., & Musdholifah, A. (2020). Enhancement of conformational B-cell epitope prediction using CluSMOTE. *Peer J. Computer Science*, 6, 1–17. 10.7717/peerj-cs.275
- Sujaree, K. (2019). Designing a vaccine cold chain network in northern Thailand using the max-min ant system. *Journal of Science and Technology*, 26(3), 257–265.
- Svanström, H., Callreus, T., & Hviid, A. (2010). Temporal data mining for adverse events following immunization in nationwide Danish healthcare databases. *Drug Safety*, 33(11), 1015–1025. doi:10.2165/11537630-000000000-00000 PMID:20925439
- Tavoschi, L., Quattrone, F., D'Andrea, E., Ducange, P., Vabanesi, M., Marcelloni, F., & Lopalco, P. L. (2020). Twitter as a sentinel tool to monitor public opinion on vaccination: An opinion mining analysis from September 2016 to August 2017 in Italy. *Human Vaccines & Immunotherapeutics*, 16(5), 1062–1069. doi:10.1080/21645515.2020.1714311 PMID:32118519
- Tomeny, T. S., Vargo, C. J., & El-Toukhy, S. (2017). Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009–15. *Social Science & Medicine*, 191, 168–175. doi:10.1016/j.socscimed.2017.08.041 PMID:28926775
- Tomic, A., Tomic, I., Dekker, C. L., Maecker, H. T., & Davis, M. M. (2019). The FluPRINT dataset, a multidimensional analysis of the influenza vaccine imprint on the immune system. *Scientific Data*, 6(1), 1–11. doi:10.1038/s41597-019-0213-4 PMID:31636302
- Tomic, A., Tomic, I., Rosenberg-Hasson, Y., Dekker, C. L., Maecker, H. T., & Davis, M. M. (2019). SIMON, an automated machine learning system, reveals immune signatures of influenza vaccine responses. *Journal of Immunology (Baltimore, Md.: 1950)*, 203(3), 749–759. doi:10.4049/jimmunol.1900033 PMID:31201239
- Trifiro, G., Sultana, J., & Bate, A. (2018). From Big Data to smart data for pharmacovigilance: The role of healthcare databases and other emerging sources. *Drug Safety*, 41(2), 143–149. doi:10.1007/s40264-017-0592-4 PMID:28840504
- Tumiotto, C., Riviere, L., Bellacave, P., Recordon-Pinson, P., Vilain-Parce, A., Guidicelli, G.-L., & Fleury, H. (2017). Sanger and next-generation sequencing data for characterization of CTL epitopes in archived HIV-1 proviral DNA. *PLoS One*, 12(9), 1–10. doi:10.1371/journal.pone.0185211 PMID:28934310

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2 • July-December 2021


- Vaishya, R., Javaid, M., Khan, I. H., & Maleem, A. (2020). Artificial intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome*, 14(4), 337–339. doi:10.1016/j.dsx.2020.04.012 PMID:32305024
- Vang, Y. S., & Xie, X. (2017). HLA class I binding prediction via convolutional neural networks. *Bioinformatics (Oxford, England)*, 33(17), 2658–2665. doi:10.1093/bioinformatics/btx264 PMID:28444127
- Wang, S. V., Maro, J. C., Baro, E., Izem, R., Dashevsky, I., Rogers, J. R., Nguyen, M., Gagne, J. J., Patorno, E., Huybrechts, K. F., Major, J. M., Zhou, E., Reidy, M., Cosgrove, A., Schneeweiss, S., & Kulldorff, M. (2018, November). Data mining for adverse drug events with a propensity score-matched tree-based scan statistic. *Epidemiology (Cambridge, Mass.)*, 29(6), 895–903. doi:10.1097/EDE.0000000000000907 PMID:30074538
- Woo, E. J., Ball, R., Burwen, D. R., & Braun, M. M. (2008). Effects of stratification on data mining in the US Vaccine Adverse Effect Reporting System (VAERS). *Drug Safety*, 31(8), 667–674. doi:10.2165/00002018-200831080-00003 PMID:18636785
- Wu, H., Mu, J., Da, T., Xu, M., Taylor, R. H., Iordachita, I., & Chirikjian, G. S. (2019). Multi-Mosquito object detection and 2D pose estimation for automation of PfSPZ malaria vaccine production. IEEE 15th International Conference on Automation Science and Engineering (CASE).
- Yildirim, M., Keskinocak, P., Pelton, S., Pickering, L., & Yildirim, I. (2020). Who is at risk of 13-valent conjugated pneumococcal vaccine failure? *Vaccine*, 38, 1671–1677. doi:10.1016/j.vaccine.2019.10.060
- Yong, B., Shen, J., Liu, X., Li, F., Chen, H., & Zhou, Q. (2020). An intelligent blockchain-based system for safe vaccine supply and supervision. *International Journal of Information Management*, 52, 1–12. doi:10.1016/j.ijinfomgt.2019.10.009
- Young, J. A., Fivelman, Q. L., Blair, P. L., de la Vega, P., Le Roch, K. G., Zhou, Y., Carucci, D. J., Baker, D. A., & Winzeler, E. A. (2005). The *Plasmodium falciparum* sexual development transcriptome: A microarray analysis using ontology-based pattern identification. *Molecular and Biochemical Parasitology*, 143(1), 67–79. doi:10.1016/j.molbiopara.2005.05.007 PMID:16005087
- Yu, H., Wang, G., Zhao, Z., Wang, H., & Wang, Z. (2019). Chicken embryo fertility detection based on PPG and convolutional neural network. *Infrared Physics & Technology*, 103, 1–6. doi:10.1016/j.infrared.2019.103075
- Yuan, X., Schuchard, R. J., & Crooks, A. T. (2019). Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. *Social Media + Society*, 5(3), 1–12. doi:10.1177/2056305119865465
- Zame, W. R., Bica, I., Shen, C., Curth, A., Lee, H.-S., Bailey, S., Weatherall, J., Wright, D., Bretz, F., & van der Schaar, M. (2020). Machine learning for clinical trials in the era of COVID-19. *Statistics in Biopharmaceutical Research*, 12(4), 506–517. doi:10.1080/19466315.2020.1797867
- Zhang, G., Lin, H. H., Keskin, D. B., Reinherz, E. L., & Brusica, V. (2011). Dana-Farber repository for machine learning in immunology. *Journal of Immunological Methods*, 374(1-2), 18–25. doi:10.1016/j.jim.2011.07.007 PMID:21782820
- Zhang, W., & Niu, Y. (2010). Predicting flexible length linear B-cell epitopes using pairwise sequence similarity. *3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010)*.
- Zhang, Y., Aevermann, B., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., Li, X., Macken, C., Mahaffey, C., Pickett, B. E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., & Scheuermann, R. H. et al. (2017). Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1), D466–D474. doi:10.1093/nar/gkw857 PMID:27679478
- Zung, J. (2019). The growing importance of real-time access to clinical study performance. *Applied Clinical Trials*, 28(11), 7.

# Predicting Inpatient Status for the Next 30/60/90 Days With Machine Learning

Lakshmi Prayaga, University of West Florida, USA

Krishna Devulapalli, Indian Institute of Chemical Technology, India

Chandra Prayaga, University of West Florida, USA

 <https://orcid.org/0000-0002-7534-4313>

Joe Carloni, Lakeview Center Inc., USA

## ABSTRACT

In this paper, the authors report the development of machine learning techniques that can help hospital authorities assess a patients' medical condition and also calculate the probability of readmission of the patient as inpatient, and thus identify patients with higher risks for readmissions. Factor analysis is performed on patient data to understand the severity of mental health, and random forest models are used to determine the probability of a patient becoming an inpatient for the next 30/60/90 days from their last visit to the physician's office. The random forest model fits the data with an overall OOB error rate of 3.69% and an accuracy of 97.65%. The accuracy on the test data was 96.11%. A web application is also developed to provide a user-friendly interface for physicians and administrators to interact with and obtain relevant information for a given patient and or a group of patients. The web application affords physicians additional inputs to assist in their diagnosis and administrators a window into anticipating and preparing for future patient needs.

## KEYWORDS

30/60/90 Day Predictions, In-Patient Stay Predictions, Interactive Web App for Predictive Analytics, Machine Learning, Mental Health Severity Index

## INTRODUCTION

Mental health illnesses are becoming more prevalent (Owens et al., 2019) in the United States. In 2019, NIH estimates that approximately one in five people or 51.5 million people aged 18 years and over suffered from mental and/or substance abuse disorders (MSUDs). Of these adults, nearly 45 million had a mental disorder alone, 11 million had a substance abuse disorder alone, and 8 million had both a mental disorder and a substance abuse disorder. It is further found that disorders such as depression, anxiety, and substance abuse are associated with significant distress and impairment, including complications with multiple chronic conditions, disability, inability to function in society, and substantial economic costs. Spornova et al. (2019) cite a 3-year adjusted mean cost at \$38,250 for those with a mental health disorder, and \$22,280 for those without a mental health disorder. According to The American Psychological Association (Winerman, 2017), in the year 2013, \$187.8 billion dollars, including out of pocket expenses, were categorized as costs related to mental disorders.

DOI: 10.4018/IJBDAH.20210701.oa9

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Taking into account additional costs associated with loss of productivity and disability payments, the total cost of MSUDs to society is estimated to be more than twice that amount.

Hospitalization is a very important component of treatment plans for individuals with serious and persistent illness. However, hospital inpatient stay has become very expensive in countries like the USA. According to Lan Liang et. al. (2016), there were over 35 million hospital stays, equating to 104.2 stays per 100,000 population. The average cost per hospital stay was \$11,700, making hospitalization one of the most expensive types of healthcare services.

According to The Piper Report (2020), hospital lengths of stay for mental health (MH) or substance abuse (SA) disorders also vary considerably, especially for mental-health related admissions. Nationwide, the MH average length of stay is 8.0 days. According to the same Report, MH and SA hospitalizations are, on average, less expensive than non-MHSA stays:

\$5,700 per MH stay.

\$4,600 per SA stay.

\$9,300 per stay for all other conditions.

Health Catalyst, in its Newsletter issue May 25, 2017 published an article entitled “Enhancing Mental Health Care Transitions Reduces Unnecessary Costly Readmissions” and stated that “Nationally, hospitalization for persons with mental health disorders has increased faster than hospitalization for any other condition”. Also mentioned is the lack of bed space to admit the patients on a timely basis.

Therefore, it becomes necessary to assess the mental health condition of the patients. In the current study, machine learning techniques are developed, for associating the patients’ demographic, behavioral, psychological and other related data, and to evaluate the probability of hospital inpatient admission for these patients. By setting a threshold value for the probability, the medical practitioner can assess whether the patient needs inpatient admission or not. It is also interesting to assess the level of Mental Health Severity of communities, based on race, gender and patient status by using all the complex and rich data that is available. Factor analysis techniques are used here to develop a comprehensive Mental Health Severity Index (MHSI) by using the variables and to rank the communities. The rest of the article is organized in the following sections: Literature review, Materials and methods, Machine learning algorithms used, Results, and Conclusion.

## LITERATURE REVIEW

Hyunyoung Baek et al. (2018) applied statistical data mining approaches to analyze the length of hospital stay using electronic health records. They identified five significant variables (frequency of surgery, frequency of diagnosis, frequency of patient transfer, severity, and insurance type) as most relevant to predict the length of stay (LOS). Multiple regression analysis was used for prediction, with an  $R^2$  of 0.267 and a mean absolute error 4.68. Luc Jansen et al. (2018) studied the extent to which medical-psychiatric comorbidities relate to health-economic outcomes in general and in different subgroups. Their study indicated that comorbidities such as depression increased the LOS for patients by 4.38 days compared to those patients who did not have comorbidities. Tsai et al. (2016) applied Artificial Neural Network (ANN) models to predict LOS for cardiac patients with coronary atherosclerosis (CAS), acute myocardial infarction (AMI), and Heart Failure (HF). Their study obtained accuracy levels of 88.07% to 89.95% for CAS patients at predischage level, and 88.31% to 91.53% at the pre-admission stage. Results for AMI and HF were observed with accuracy levels of 64.12% to 66.78% at pre discharge level and at 63.69% to 67.47% at the preadmission state. Mekhaldi et al. (2020) applied the Random Forest and Gradient Boosting models to predict length of stay (LOS) in a hospital setting. Lin et al. (2016) used multivariate logistic regression model to predict inpatient readmission and outpatient admission in the elderly, and expressed that these models provide a basis for wider application in National health Service. They predicted Length of Hospital Stay for older citizens considering comorbidities, home healthcare, and prior use of healthcare facilities. They applied

the Area Under the Curve (AUC) model to determine the LOS and reported the AUC of the inpatient readmission model as 0.655. Gopalakrishna, Ithman and Malwitz (2015) studied the predictors of length of stay in acute psychiatric hospitals. They applied regression models with natural logarithms of LOS as the dependent variable and age, marital status, involuntary admission and diagnosis of an affective disorder or a psychotic disorder as independent variables, which could explain about 20% of the variation of the variance of LOS.

Hospital readmissions have received attention from researchers, in view of the cost of such readmissions. Upadhyay, Stephenson and Smith (2019) studied the effect of readmission rates on hospital financial performance. Their analysis of data, from 98 hospitals in the State of Washington from 2012 to 2014, indicated that a reduction in acute myocardial infarction (AMI) readmission rates is related with increased operating revenues as expenses associated with costly treatments related with unnecessary readmissions are avoided. Cardarelli et al (2018) carried out a quasi-experimental study design which assessed implementation of a lay health worker (LHW) model for assisting high-risk patients with their post-discharge social needs. The LHW intervention involved assessment and development of a personalized social needs plan for enrolled patients (e.g. transportation and community resource identification), with post-discharge follow-up calls, and resulted in a 47.7% relative reduction of 30-day hospital readmissions rates between baseline and intervention phases of the study. Wan et al (2011) have given an extensive review of preventable hospital readmissions. Liu et al (2020) used artificial neural networks (ANN) to predict 30-day hospital readmissions. They compared the performance of their models with hierarchical logistic regression models and found that the ANNs increased the AUC for prediction of 30-day readmissions. Flaks-Manov, N., Topaz, M., Hoshen, M. *et al.* (2019) suggest that readmission risk identification should incorporate a two time-point approach in which preadmission data is used to identify high-risk patients as early as possible during the index admission and an “all-hospital” model is applied at discharge to identify those that incur risk during the hospital stay.

In the field of mental health, Šprah, L., Dernovšek, M.Z., Wahlbeck, K. *et al.* (2017) have reviewed the impact of physical comorbidity variables on readmission after discharge from psychiatric or general inpatient care among patients with co-occurring psychiatric and medical conditions. Benjenk and Chen (2018) have reviewed Effective mental health interventions to reduce hospital readmission rates.

Researchers (Degenhardt et al., 2019; Wongvibulsin, Wu and Zeger 2020) suggest that Random Forest (RF) algorithms are good candidates to address the challenges associated with high dimensional and heterogeneous data that includes electronic health records. Degenhardt et al. report that RF methods have been applied in proteomics. RF has been successfully applied in genetics, gene expression, methylation, proteomics, and metabolomics studies. It is a flexible approach that can be used to perform both classifications, i.e., predicting case-control status, and regression, i.e., predicting quantitative traits. Wongvibulsin, Wu and Zeger (2020) also used the RF algorithm to predict sudden cardiac arrests with a high degree of confidence. Based on the included literature review, we find that our choice of RF is suitable for this study since a. it is a good technique to use for high dimensional data as reported by (Degenhardt et al., 2019; Wongvibulsin et al., 2019) and the data for the current study falls under this category, b. the 96.31% accuracy obtained by using RF in the current study was higher than those reported from earlier studies using regression analysis (Lin et. al 2016) and neural networks (Tsai et al. 2016) produced accuracy levels of 88.07% to 89.95% and c. it has a wide applicability as reported in the literature review section.

Our contribution to the literature on predicting inpatient stay is that we address the probability of inpatient admission for patients with mental health illnesses using only demographic and psychosocial data and not requiring clinical data. Additionally, we use two machine learning algorithms, one, Factor Analysis to determine the severity of the mental illness for specific population groups of the dataset and two, Random Forest to predict the probability of a given patient becoming an inpatient in the next 30, 60 or 90 days. We also develop a web application for physicians and administrators to search for a specific patient and obtain the probability of that patient becoming an inpatient. The

web application also provides group wise information for a specific population from the available dataset. It is a useful tool to assist physicians to get an additional input to their diagnosis and can be used to prepare a treatment plan for that patient. It also allows administrators to use the tool to plan for future resources required using the 30-/60-/90-day search options leading to better patient care.

## **MATERIALS AND METHODS**

### **Data Description**

The dataset used in this study was provided by Lakeview Center, Inc., which is a private non-profit organization providing behavioral health care. The data was anonymized by Lakeview Center and then provided for purposes of this study, thus no personal information was compromised, maintaining strict confidentiality and ethical norms. The data is available in an Excel file and consists of 80849 observations. The Excel file contains the data relating to daily admissions of patients during the period January 2019 to June 2020. This data contains information on 20524 patients with different mental disorders. Among the 20524 patients, 2754 (13.42%) patients are inpatients and the remaining 17770 (86.58%) are outpatients.

Data is collected broadly utilizing three types of forms, viz., psychosocial, service needs assessment and SAFET assessment. Psychosocial data relates to items such as Symptoms, Onset, Frequency, Severity, Use of Medicines, Previous mental health treatment, and Family history of Mental Illness. Service Needs assessment is used to identify strengths and needs of individuals that may impact their ability to participate in services. Based on that assessment, the medical practitioner would provide case management services to reduce such barriers. SAFET is the instrument used to identify homicide & suicide risk that is performed on all clients over the age of 5 years.

Specimen screenshots of each of the three categories are displayed in Figures 1 to 3.

Each observation contains data on 145 variables, relating to ClientID, activityMonth, ZipArea and diagnosis in the categories of Psychosocial assessment, Service Needs Assessment and SAFET. Out of these, 135 variables are retained for the current study. The above-mentioned client specific variables are not used in the model fitting, and seven other variables, with only one value entered, are also eliminated.

All the variables contain either character values or numerical values in a scale of 0 to 10. All the missing values are filled with the character string 'Unknown'. Demographic variables such as race, sex, and ethnicity contain character data. Variables with values from 0 to 10 denote the ranking of the diagnostic test result. Higher ratings represent higher levels of severity of mental illness. The variable numinpatientStayLast30 represents the number of times the patient was admitted in the hospital as an inpatient during the last 30 days with reference to activityMonth. This variable contains values ranging from 0 to the number of times the patient was admitted in the hospital during the last 30 days. For patients not admitted in the hospital, these fields are filled with the value zero, and for those who were admitted, they are assigned the values 1, 2, 3 ... corresponding to the number of times they were admitted.

### **Data Preprocessing**

All the character data is recoded with numerical values. For example, the variable sex is recoded by numerical values 1 and 2, 1 for males and 2 for females. Similarly, the other demographic variables are also recoded with appropriate numerical values. All the missing values are assigned the value 9. In the case of all the other remaining categorical variables having character data, appropriate numerical values are assigned. For example, if any variable has character values 'yes', 'no', 'unknown', recoding is done by assigning the numerical values 1, 2 and 9. Variables having numerical values are not modified.

The variable numInpatientStayLast30 contains the number of times the patient was admitted in the hospital during the Last 30 days and it takes any value from 0 to any number. This variable is converted

Figure 1. Specimen data entry screenshot of psychosocial data

into a binary variable, representing the patient status as InPatient or OutPatient, as follows : A patient is considered as InPatient if he/she has stayed at least once for one or more days and Outpatient if the patient has not stayed at least once. Accordingly, if the variable numInpatientStayLast30 takes any value greater than 0, it is converted to 1 (representing Present) and to 0 if its value is zero (representing Absent). Thus, the value 1 represents the InPatient status as present (ie. the patient is an InPatient), and zero represents the InPatient status as absent (ie. the patient is an OutPatient). This variable is treated as the dependent variable and all the remaining variables are treated as independent variables in fitting the Random Forest model.

## MACHINE LEARNING ALGORITHMS USED FOR THE STUDY

### Factor Analysis for evaluating Mental Health Severity Index (MHSI)

Using the variables under consideration, a comprehensive Mental Health Severity Index (MHSI) is calculated by the procedure detailed in Prayaga et al. (2020). Factor analysis is a dimension reduction technique to reduce a large number of variables into a fewer number called factors. In this study, we have used principal component analysis to extract the factors. Factor analysis evaluates three important quantities viz., factor loadings, eigenvalues and factor scores. Factor loadings are essentially the correlation between the original variables and the factors. Eigenvalues show the variance explained by each factor out of the total variance. Factor scores  $F_j$ 's are index variables obtained as optimally-weighted linear combinations of the variables.

Figure 2. Specimen data entry screenshot of Service Needs Assessment

The first step in factor analysis is to determine the number of factors to be retained for further analysis. The eigenvalues are good indicators for determining the number of factors. Generally, the first few factors have eigenvalues greater than one. If the eigenvalue is greater than one, that factor should be included, else, it should be discarded. The Scree plot proposed by Ledesma et al. (2015) has also been used by researchers to assess graphically the number of factors ‘m’ to be retained for exploratory factor analysis. A Scree plot is a line plot of the eigenvalues of factors. In the Scree diagram, the number of factors ‘m’ to be retained is obtained as the meeting point of the eigenvalues curve and the parallel analysis curve.

The Mental health severity index ( $MHSI_k$ ) for each patient  $k$  is obtained by multiplying the square of each retained factor score  $F_j$  by the proportion of variance  $S_j$  explained by the corresponding factor as the weight, and then adding the products as given by the following formula:

$$MHSI_k = \sum_{j=1}^m F_j^2 S_j \quad (1)$$

where  $k = 1 \dots n$  is the number of patients and  $j = 1 \dots m$  is the number of factors

The aggregated mean MHSI values were evaluated for the following combinations of groups:

Race (4 in number) – White, Black, Others (Multiracial, American Indian etc.) and Unknown

Gender (2 in number) – Male, Female

Patient status (2 in number) – Inpatient, Outpatient

This yields 16 mean values corresponding to all possible combinations of 4 races, 2 gender classifications and 2 patient status categories ( $4 \times 2 \times 2$ ). These final mean MHSI values were then used to compare the mental health status among these 16 groups.

## Probability Of Admission By Applying Random Forest Models

A random forest (RF) is an ensemble bagging or averaging method that aims to reduce the variance of individual trees by randomly selecting (and thus de-correlating) many trees from the dataset, and



Figure 3. Specimen data entry screenshot of SAFET

averaging them. It is an extension of bagging. Random forest achieves better accuracy by reducing variance through the averaging of the prediction of orthogonal trees. It is an ensemble modeling technique that combines several machine learning algorithms into one prediction model. Research suggests that RFs improve accuracy by reducing the estimator variance by a factor of three-fourths (Genuer, 2012). Several recent studies (Blankers et al., 2020), have demonstrated that RFs have been very effective in predicting the desired outcomes with a high degree of accuracy. It is for these reasons, of reduction in the variance and improved accuracy for high dimensional data, that the Random Forest algorithm is chosen in this study.

A Random Forest Model is applied to the cleaned, preprocessed data by considering the numinpatientStayLast30 as the dependent variable and all the other variables as independent variables.

An interactive Shiny App is also developed to display the various results of the application including the model fitting, its accuracy, the important variables, patients with highest probability of admission etc. The Shiny App can also predict the probability of admission for any single patient, even in the case of new patients. The development of the Shiny App is also carried out using RStudio and is uploaded on shinyapps.io website.

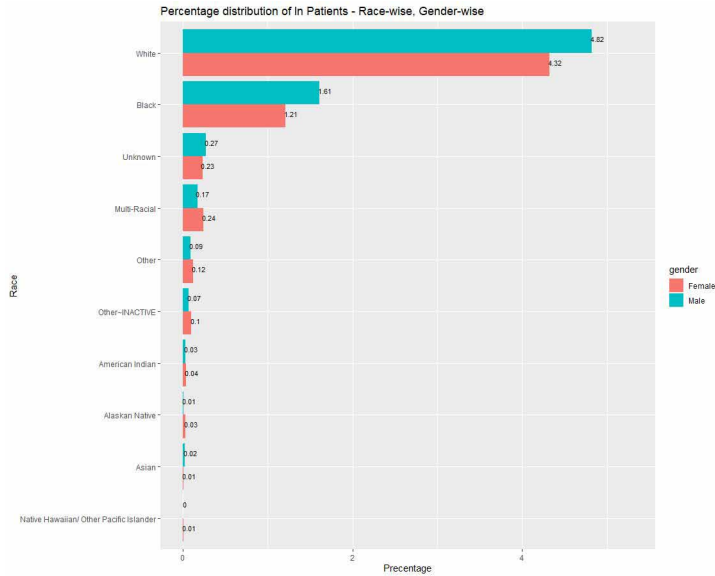
All the statistical analysis, calculation of the MHSI by factor analysis and random forest model fitting were carried out using R and associated statistical packages.

## Results and Discussion

The percentage distribution of inpatients, by Race and Gender is evaluated and displayed in Figure 4.

In order to compare the mental health severity among the races, genders and patient status (InPatient and OutPatient), factor analysis was applied to the data. The six largest eigenvalues obtained from factor analysis were 21.653, 3.432, 1.312, 1.129, 1.088 and 1.023, which are all greater than 1. The corresponding factors were therefore retained for further analysis. The Scree plot technique is also used to determine the number of factors to retain, which explain maximum variation in the data. Figure 5 shows the Scree plot. It is seen in this figure that the two curves of eigenvalues and

Figure 4. Percentage distribution of InPatients by Race and Gender with numinpatientStay30 = 1.



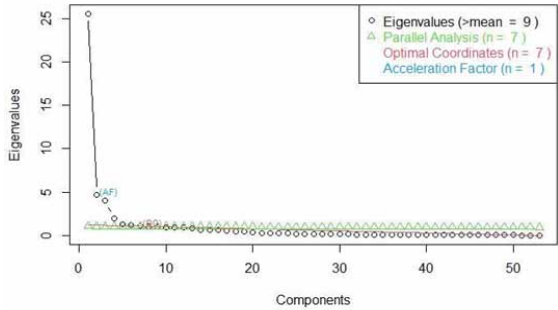
parallel analysis meet at number of components (or Factors) equal to six, suggesting that six factors be retained. As both the eigenvalues and the scree plot have identified the number of factors as 6, it was decided to proceed with the first six factors for further analysis.

The results of the proportion of variance explained by each factor and the cumulative variance are given in Table 1. As seen in the table, the cumulative variance explained by these first six factors is as high as 69.2% of the total variance.

In order to assess the mental health status among communities, the following three groups are considered:

- Race: Black, White, Others, Unknown
- Gender: Male, Female
- Patient Status: inpatient, Outpatient

Figure 5. Scree diagram to identify the number of factors to be retained.



**Table 1. Proportion of variance explained by the six factors**

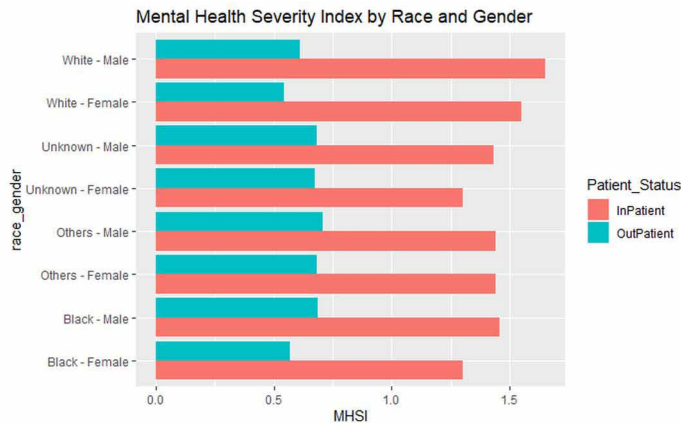
Description	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
SS loadings	19.347	6.613	6.067	3.103	0.829	0.703
Proportion Var	0.365	0.125	0.114	0.059	0.016	0.013
Cumulative Var	0.365	0.490	0.604	0.663	0.678	0.692

MHSI scores are evaluated for each patient by using formula (1) given in the Machine Learning Algorithms section above. From these individual MHSI scores, the cross tabulated mean scores or Mental health severity index(MHSI) scores are evaluated for the above three groups ie. race, gender and patient status. These results are presented in Table 2 and a bar plot in Figure 6. From the bar plot of Figure 6, it is observed that in general, MHSI scores for inpatients are higher than for outpatients in all races and genders. This clearly shows the distinction between the mental health severity for inpatients and outpatients.

**Table 2. MHS Index race-wise, gender-wise and patient-status-wise**

Patient status	Gender	Race	MHSI Score
Outpatient	Male	Unknown	0.7155423
Outpatient	Male	Others	0.7728059
Outpatient	Male	Black	0.7266442
Outpatient	Male	White	0.6436330
Outpatient	Female	Unknown	0.6881292
Outpatient	Female	Others	0.7507318
Outpatient	Female	Black	0.5957075
Outpatient	Female	White	0.5739292
inpatient	Male	Unknown	1.1698302
inpatient	Male	Others	1.2524319
inpatient	Male	Black	1.3929018
inpatient	Male	White	1.6482094
inpatient	Female	Unknown	0.9796764
inpatient	Female	Others	1.2876218
inpatient	Female	Black	1.2092737
inpatient	Female	White	1.4706496

Figure 6. Bar Plot of MHSI Scores for different races, genders and patient status



To assess the mental health severity among the inPatients, a bar plot of MHSI scores is shown for inPatients only in Figure 7, which displays the MHSI scores of inPatients belonging to different races and genders. From this plot, it is observed that in each race among inPatients, the MHSI scores for males are slightly higher than those for females, except in the case of the Others category. It is also observed that among the races and genders, white males and white females appear to have slightly higher mental health severity scores than all other categories of inPatients.

## Random Forest Model Results

A Random Forest model is fitted to the training dataset to evaluate the probability of admission as InPatient. Results of the fitted model for the training dataset are presented in Table 3. These results contain the confusion matrix and the Out Of the Bag(OOB) estimate of error rate. The Random Forest model fitted the training data very well with an overall OOB Error rate as low as 3.69% and an accuracy of 96.31%.

The fitted model is then applied to the test dataset and the confusion matrix is generated to test the accuracy of the model for the test data. These results are presented in Table 4. In the test dataset also, the accuracy is 96.11%, as can be seen from these results.

The Random Forest model also identified the variables of importance based on the meanDecreaseGini criterion (Breiman, 2001). Figure 8 shows the top ten variables identified by the Random forest model based on this criterion. From this plot, it is observed that the variables mhIssueSeverity, hasDrugUseAterWalking, hasCriticizedrug, doesWantReduceDrug etc. are the most important variables for classification and for evaluating the probabilities. Most of these important variables relate to the severity of mental health and drug abuse and the study has highlighted the fact that drug related variables contribute more towards the severity of mental health.

Figure 7. Bar Plot of MHSI among inpatients – races and sexes

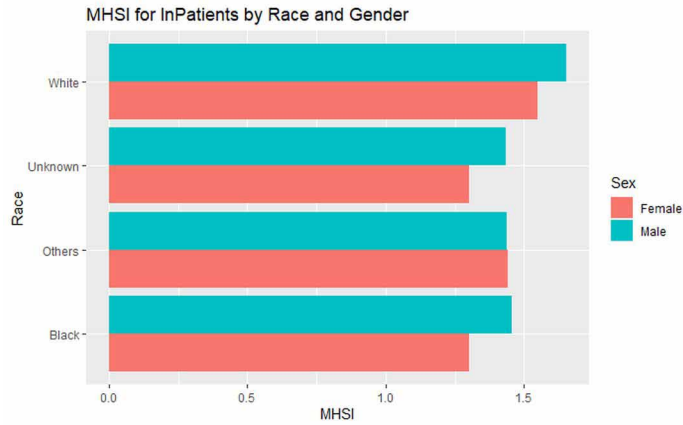


Figure 7a.

```
Call:
  randomForest(formula = traindata$numInpatientInLast30 ~ ., data = traindata)
  Type of random forest: classification
    Number of trees: 500
No. of variables tried at each split: 11

  OOB estimate of error rate: 3.69%
Confusion matrix:
      0      1 class.error
0 56470  542 0.009506771
1  1844 5824 0.240479917
```

Figures 7b.

```
Confusion Matrix and Statistics

      Reference
Prediction  0      1
 0  14115   138
 1    491  1425

      Accuracy : 0.9611
      95% CI   : (0.958, 0.964)
  No Information Rate : 0.9033
   P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 0.7977

  Mcnemar's Test P-Value : < 2.2e-16

      Sensitivity : 0.9664
      Specificity : 0.9117
   Pos Pred Value : 0.9903
   Neg Pred Value : 0.7437
      Prevalence : 0.9033
   Detection Rate : 0.8730
  Detection Prevalence : 0.8815
   Balanced Accuracy : 0.9390

      'Positive' Class : 0
```

Figure 8. Importance of Variables graph by Random Forest Model

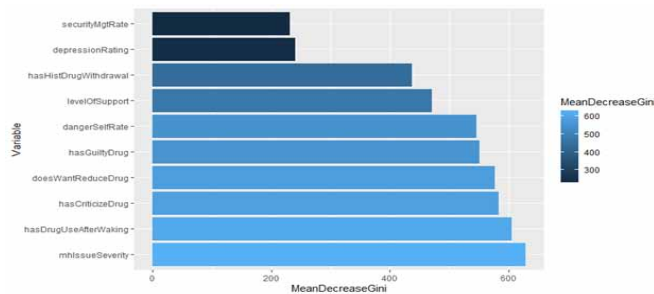
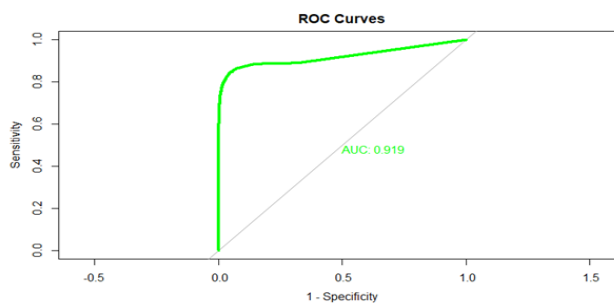


Figure 9. ROC Curve for the Random Forest Model



To evaluate the performance of the Random Forest model, the Receiver Operating Characteristic (ROC) curve is developed, and the Area under the Curve (AUC) is evaluated (Kun-Pie Lin et al., 2016). These results are displayed in Figure 9. From this plot, it is observed that the AUC value = 0.919, which is quite high, indicating that the Random Forest model has separated the two categories inpatient and outpatient very well, and the accuracy of the model is quite high.

### Shiny App for Probability Predictions

A Shiny app is developed to interactively view the Random Forest model results and to predict the probabilities of admission of the existing clients and new Clients. The shiny app is hosted on shinyapps.io website and has eight menu options as shown in Figure 10.

The first Menu “About” of the app displays information about the details of the shiny app as shown in Figure 10. The second Menu “Data Details” gives details of the dimensions of the data such as number of observations, number of variables, storage requirements, etc. The third menu option, “Explore Data,” displays the first few records of the dataset utilized. The fourth Menu option, “R.F.Model” displays the results of the Random Forest model, the OOB estimation of Error rate, the Confusion Matrix, etc., as shown in Figure 11. As seen in this plot, the model has fitted the data very accurately with an overall OOB estimation of an Error rate of 3.51%.

Figure 10. Display of the Main Menu of the Shiny App

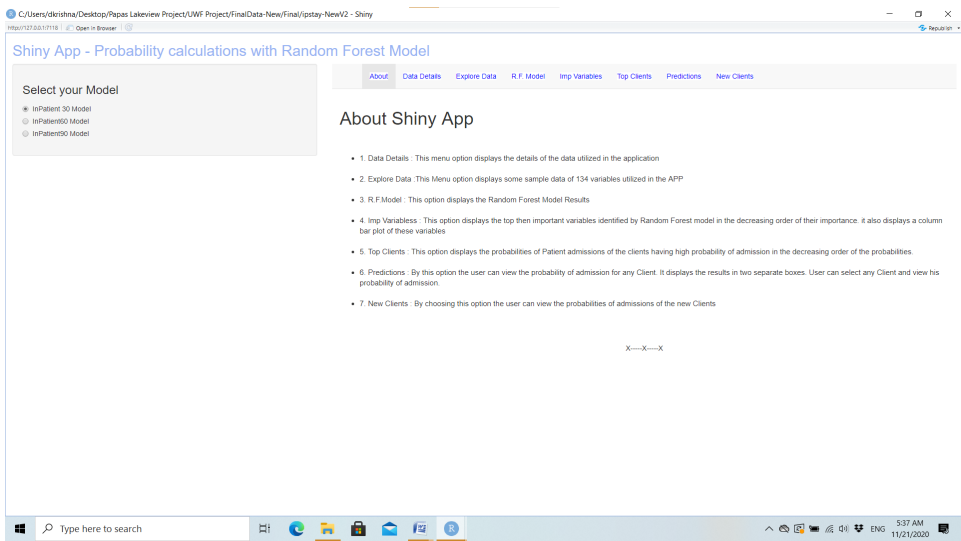
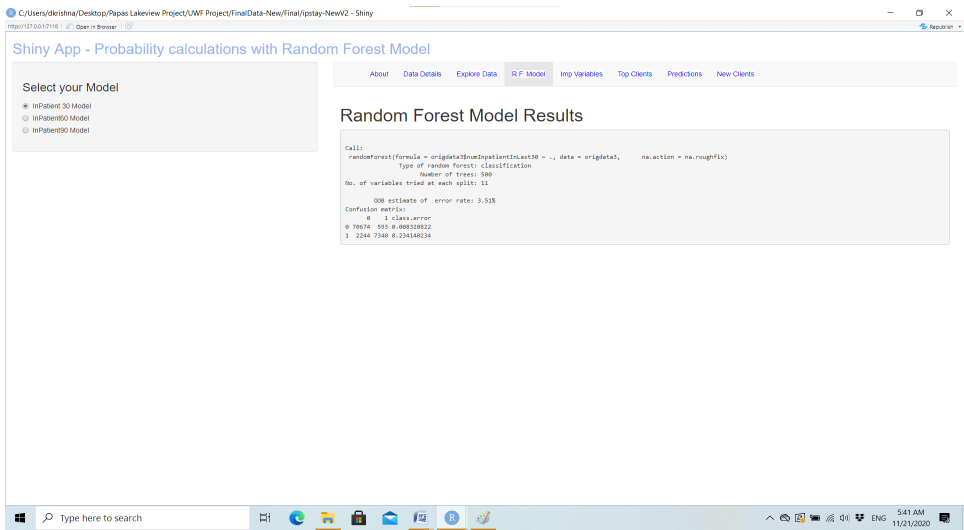


Figure 11. Display of Shiny app Random Forest model Results



The fifth Menu, “Imp Variables,” displays the top 10 important variables identified by the Random Forest Model. The sixth Menu, “Top Clients,” displays the top 1000 clients with the highest probabilities of inpatient admission. It shows a table consisting of the clientID column and the column of the corresponding probability of inpatient admissions.

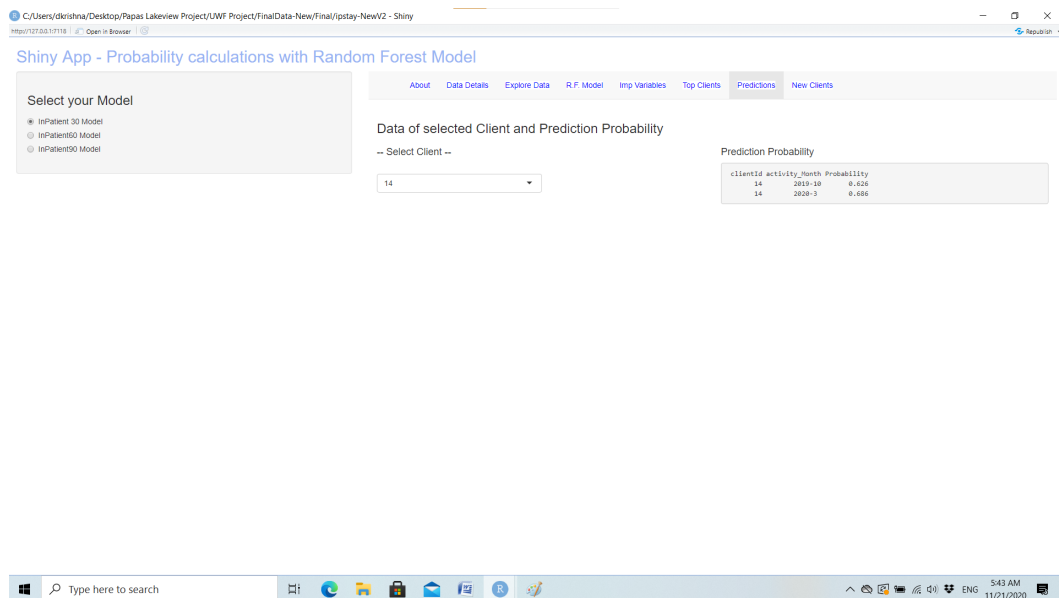
The seventh Menu option, “Predictions,” displays an interactive screen, as shown in Figure 12. On the left side of the Menu, the user can select one of the options: inpatient30 Model / inpatient

60 Model/ inpatient 90 Model for 30/60/90 days stay from the “Select your Model” radio buttons. A list box “Select Client” is provided to select any clientID from the existing clients for whom the probabilities of inpatient admissions are required. Once the clientID is selected, the text box “Prediction Probabilities” is displayed, consisting of the probabilities of inpatient admissions. This text box displays a table with the results clientID, activityMonth and the corresponding probability. As the input data consists of various admission records of the same patient at different dates, denoted by activityMonth, this table displays the probabilities for all these different dates of admission.

The user can select any other model and any clientID, and the Shiny App automatically recalculates the probabilities for the selected options and displays the results. From the plot of Figure 12, one can see that for the client with clientID 14, the prediction probability for activity-month 2019-10 is 0.626, and for the activity month 2020-3 it is 0.686.

The eighth Menu option, “New Clients,” allows the user to upload the data of new patients and view the probability of that patient becoming an inpatient. .

**Figure 12. Display of Predictions screen of shiny app**



## Conclusions

An attempt is made in this study to evaluate the probability of patient admission as inpatient utilizing the data collected from Lakeview Center Inc. The distribution of patients with mental health disorders is evaluated race-wise, gender-wise and patient-status-wise. A novel Mental Health Severity Index (MHSI) is developed using factor analysis. Utilizing the MHSI scores, the mental health status of the patients is studied for the categories of race, gender, and patient status. It was found that both male and female Caucasian patients appear to have slightly higher mental health severity index compared to other races. The mental health severity of InPatients is higher than the OutPatients.

The machine learning technique RandomForest (RF) model is applied to the patients data to assess the probability of readmission. The Random Forest Model has identified the variables mhlIssueSeverity (mental health severity) and other drug abuse related variables as important variables for classification. It appears that drug abuse related variables play an important role in mental health severity. The RF



model could accurately classify the patients with an overall OOB estimate of Error rate of 3.51%. The accuracy of the model is tested by various metrics including confusion matrix, OOB Error rate, ROC Curve and Area Under the Curve (AUC) and all the metrics have yielded high values indicating the accuracy of the fitted model. An interactive shiny app is also deployed on shinyapps.io website to display the results of the Factor Analysis model and the results of Random Forest model.

A limitation of the study is the unavailability of clinical data; we plan to acquire clinical data and study the impact of comorbidities on readmissions of patients with mental illnesses.

## REFERENCES

- Alam, Z., Rahman, S., & Rahman, M. S. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, 15, 100180. doi:10.1016/j.imu.2019.100180
- Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., & Yoo, S. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLoS One*, 13(4), e0195901. doi:10.1371/journal.pone.0195901 PMID:29652932
- Benjenk, I., & Chen, J. (2018). Effective mental health interventions to reduce hospital readmission rates: A systematic review. *Journal of Hospital Management and Health Policy*, 2, 45. doi:10.21037/jhmhp.2018.08.05 PMID:30283917
- Blankers, M., van der Post, L., & Dekker, J. (2020). Predicting hospitalization following psychiatric crisis care using machine learning. *BMC Medical Informatics and Decision Making*, 20(1), 332. doi:10.1186/s12911-020-01361-1 PMID:33302948
- Breiman, L. (2001). Random Forests. *J Mach Learn*, 45(1), 5–32. doi:10.1023/A:1010933404324
- Cardarelli, R., Horsley, M., Ray, L., Maggard, N., Schilling, J., Weatherford, S., Feltner, F., & Gilliam, K. (2018, February). Reducing 30-day readmission rates in a high-risk population using a lay-health worker model in Appalachia Kentucky. *Health Education Research*, 33(1), 73–80. doi:10.1093/her/cyx064 PMID:29474535
- Cattell, R. B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276. doi:10.1207/s15327906mbr0102\_10 PMID:26828106
- Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, 20(2), 492–503. doi:10.1093/bib/bbx124 PMID:29045534
- Enhancing Mental Health Care Transitions Reduces Unnecessary Costly Readmissions. (2017). Retrieved from <https://www.healthcatalyst.com/wp-content/uploads/2017/05/Enhancing-Mental-Health-Care-Transitions-Reduces-Unnecessary-Costly-Readmissions.pdf>
- Flaks-Manov, N., Topaz, M., Hoshen, M., Balicer, R. D., & Shadmi, E. (2019). Identifying patients at highest-risk: The best timing to apply a readmission predictive model. *BMC Medical Informatics and Decision Making*, 19(1), 118. doi:10.1186/s12911-019-0836-6
- Genuer, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3), 1–20. doi:10.1080/10485252.2012.677843
- Gopalakrishna, G., Ithman, M., & Malwitz, K. (2015). Predictors of Length of Stay in an Acute Psychiatric Hospital. *International Journal of Psychiatry in Clinical Practice*, 19(4), 238–244. doi:10.3109/13651501.2015.1062522 PMID:26073671
- Jansen, L., Schijndel, M. V., Waarde, J. V., & Busschbach, J. V. (2018). Health-economic outcomes in hospital patients with medical-psychiatric comorbidity: A systematic review and meta-analysis. *PLoS One*, 13(3), e0194029. doi:10.1371/journal.pone.0194029 PMID:29534097
- Kabacoff, R. (n.d.). *Principal components and factor analysis*. Retrieved from <https://www.statmethods.net/advstats/factor.html>
- Ledesma, R. D., Valero-mora, P. M., & Macbeth, G. (2015). The Scree Test and the Number of Factors: A Dynamic Graphics Approach. *The Spanish Journal of Psychology*, 18, 18. doi:10.1017/sjp.2015.13 PMID:26055575
- Liang, L., Moore, B., & Soni, A. (2020). *National inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2017: Statistical Brief #261*. Agency for Healthcare Research and Quality.
- Lin, K. P., Chen, P. C., Huang, L. Y., Mao, H. C., & Chan, D. D. (2016). Predicting inpatient Readmission and Outpatient Admission in Elderly. *Medicine*, 95(16), e3484. doi:10.1097/MD.0000000000003484 PMID:27100455
- Liu, W., Stansbury, C., Singh, K., Ryan, A. M., Sukul, D., Mahmoudi, E., Waljee, A., Zhu, J., & Nallamotheu, B. K. (2020). Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PLoS One*, 15(4), e0221606. doi:10.1371/journal.pone.0221606 PMID:32294087

- Mekhaldi, R. N., Caulier, P., Chaabane, S., Chraibi, A., & Piechowiak, S. (2020). Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting. *Trends and Innovations in Information Systems and Technologies*, 1159, 202–211. doi:10.1007/978-3-030-45688-7\_21
- Narkhede, S. (2018). *Understanding AUC – ROC Curve*. Retrieved from: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- Owens, P. L., Fingar, K. R., McDermott, K. W., Muhuri, P. K., & Heslin, K. C. (2019). Inpatient Stays Involving Mental and Substance Use Disorders, 2016: Statistical Brief #249. Agency for Healthcare Research and Quality.
- Piper, K. (2020). *Medicare, Medicaid, health reform*. Retrieved from <http://www.piperreport.com/>
- Prayaga, L., Devulapalli, K., & Prayaga, C. (2020). Combining clustering and factor analysis as complimentary techniques. *International Journal of Data Analytics*, 1(2), 48–57. doi:10.4018/IJDA.2020070104
- Sporinova, B., Manns, B., Tonelli, M., Hemmelgarn, B., MacMaster, F., Mitchell, N., Au, F., Ma, Z., Weaver, R., & Quinn, A. (2019). Association of Mental Health Disorders With Health Care Utilization and Costs Among Adults With Chronic Disease. *JAMA Network Open*, 2(8), e199910. doi:10.1001/jamanetworkopen.2019.9910 PMID:31441939
- Šprah, L., Dernovšek, M. Z., Wahlbeck, K., & Haaramo, P. (2017). Psychiatric readmissions and their association with physical comorbidity: A systematic literature review. *BMC Psychiatry*, 17(1), 2. doi:10.1186/s12888-016-1172-3 PMID:28049441
- Tsai, P. F., Chen, P. C., Chen, Y. Y., Song, H. Y., Lin, H. M., Lin, F. M., & Huang, Q. P. (2016). Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. *Journal of Healthcare Engineering*, 2016, 7035463. doi:10.1155/2016/7035463 PMID:27195660
- Upadhyay, S., Stephenson, A. L., & Smith, D. G. (2019). Readmission Rates and Their Impact on Hospital Financial Performance: A Study of Washington Hospitals. *Inquiry*, 56. doi:10.1177/0046958019860386 PMID:31282282
- Wan, H., Zhang, L., Witz, S., Musselman, K. J., Yi, F., Mullen, C., Benneyan, J., Zayas-Castro, J., Rico, F., Cure, L., & Martinez, D. (2016). A literature review of preventable hospital readmissions: Preceding the Readmissions Reduction Act. *IIE Transactions on Healthcare Systems Engineering*, 6(4), 193–211. doi:10.1080/19488300.2016.1226210
- Winerman, L. (2017). By the numbers: The cost of treatment. *Monitor on Psychology*, 48(3).
- Wongvibulsin, S., Wu, K. C., & Zeger, S. L. (2019). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Medical Research Methodology*, 20(1), 1. doi:10.1186/s12874-019-0863-0 PMID:31888507

**International Journal of Big Data and Analytics in Healthcare**

Volume 6 • Issue 2

*Lakshmi Prayaga is a Professor in the Department of Information Technology, University of West Florida. Her research focuses on applications of technology in healthcare, sports medicine, management and training. Topics of interest include robotics, data visualizations and analytics. She has co-authored books on robotics, Android App development, beginning game programming, programming the web with ColdFusion and XHTML, and using game programming to teach computer science concepts. She has also published numerous publications in International journals and conferences. She teaches graduate and undergraduate courses in Data Analytics, Data Visualizations, Machine learning and Script programming. She has an Ed.D. in Instructional Technology and a M.S. in Software Engineering, both from UWF and an MBA from Alabama A&M University.*

*Krishna Devulapalli is a retired scientist from the Indian Institute of Chemical Technology, India. Currently he is a Freelance Data Scientist and is recognized as TapChief Expert in Data Science by TapChief, India. His research interests include applied statistics in multiple domains such as correlations among physico-chemical attributes of substances, healthcare analytics, BioInformatics, BioStatistics, Chemometrics, Reliability Studies, Pattern Recognition, Neural Networks, Rule Based Systems, Machine Learning etc. He has published more than 30 research papers in various journals and also presented number of papers in conferences and seminars. He has contributed some chapters in books related to Medical Statistics. He is a Member of various professional Societies like Indian Society of Medical Statistics, Computer Society of India, Indian Society of Analytical Scientists etc. He is recognized as a Guest Faculty in various organizations like Statistics Department, Osmania University, NIPER Guwahati, IICT, CSI, CMC, etc.*

*Chandra Sekhar Prayaga is currently Professor of Physics, University of West Florida (UWF). He has a Ph.D. in Physics from the Indian Institute of Science, Bangalore, India (1975), where he was also a faculty member from 1981 to 1987. He has 40-plus years of experience in teaching physics, and has helped raise more than \$3 million in funding for research and projects involving University of West Florida faculty and students. His current research interests include optical and electronic properties of liquid crystals, Langmuir-Blodgett films, phase transitions and laser spectroscopy, physics education and data analytics. He mentors undergraduate student research projects, and coordinates summer camps on science and technology for middle and high school students. He is cofounder of Discovery Spot: A technology playground for middle and high school students to experience the latest technologies with hands-on activities, such as building smart Cities using IoT. He is co-author of a book, "Robotics: A Project-Based Approach," Cengage Publishers (2014).*

# International Journal of Big Data and Analytics in Healthcare

Volume 6 • Issue 2 • July-December 2021 • ISSN: 2379-738X • eISSN: 2379-7371

## MISSION

The mission of the **International Journal of Big Data and Analytics in Healthcare (IJBD AH)** is to provide timely and innovative research on the ways in which big data is revolutionizing the medical and healthcare fields. This journal aims to encourage the further development of applications and practice relating to the management and analysis of large amounts of data in the healthcare sector as well as provide a framework for future research in the field.

## SUBSCRIPTION INFORMATION

The International Journal of Big Data and Analytics in Healthcare (IJBD AH) is available in print and electronic formats and offers individual or institution-level pricing. Full subscription information can be found at [www.igi-global.com/IJBD AH](http://www.igi-global.com/IJBD AH).

IJBD AH is also included in IGI Global's InfoSci-Journals Database which contains all of IGI Global's peer-reviewed journals and offers unlimited simultaneous access, full-text PDF and XML viewing, with no DRM. Subscriptions to the InfoSci-Journals Database are available for institutions. For more information, please visit [www.igi-global.com/infosci-journals](http://www.igi-global.com/infosci-journals) or contact E-Resources at [eresources@igi-global.com](mailto:eresources@igi-global.com).

## CORRESPONDENCE AND QUESTIONS

### EDITORIAL

Mu-Yen Chen, Editor-in-Chief • [IJBD AH@igi-global.com](mailto:IJBD AH@igi-global.com)

### SUBSCRIBER INFO

#### IGI Global • Customer Service

701 East Chocolate Avenue • Hershey PA 17033-1240, USA

**Telephone:** 717/533-8845 x100 • **E-Mail:** [cust@igi-global.com](mailto:cust@igi-global.com)